University of Helsinki
Faculty of Arts
Department of Digital Humanities
Language Technology

Master's Thesis

Analysing Finnish Multi-Word Expressions with Word Embeddings

Sami Itkonen
014026847

Supervisor: Jörg Tiedemann

20.4.2020

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto – Fakultet/Sektion – Faculty Humanistinen	Laitos – Institution – Department Digitaaliset ihmistieteet	
Tekijä – Författare – Author Sami Itkonen		
Työn nimi – Arbetets titel – Title Analysing Finnish Multi-Word Expressions with Word Embeddings		
Oppiaine – Läroämne – Subject Kieliteknologia		
Työn laji – Arbetets art – Level Pro gradu	Aika – Datum – Month and year 04/2020	Sivumäärä– Sidoantal – Number of pages 80

Tiivistelmä – Referat – Abstract

Sanayhdistelmät ovat useamman sanan kombinaatioita, jotka ovat jollakin tavalla jähmeitä ja/tai idiomaattisia. Tutkimuksessa tarkastellaan suomen kielen verbaalisia idiomeja sanaupotusmenetelmän (word2vec) avulla. Työn aineistona käytetään Gutenberg-projektista haettuja suomenkielisiä kirjoja.

Työssä tutkitaan pääosin erityisesti idiomeja, joissa esiintyy suomen kielen sana ‘silmä’. Niiden idiomaattisuutta mitataan komposiittisuuden (kuinka hyvin sanayhdistelmän merkitys vastaa sen komponenttien merkitysten kombinaatiota) ja jähmeyttä leksikaalisen korvaustestin avulla. Vastaavat testit tehdään myös sanojen sisäisen rakenteen huomioonottavan fastText-algoritmin avulla. Työssä on myös luotu Gutenberg-korpuksen perusteella pienehkö luokiteltu lausejoukko, jota lajitellaan neuroverkkopohjaisen luokittelijan avulla. Tämä lisäksi työssä tunnustellaan eri ominaisuuksien kuten sijamuodon vaikutusta idiomin merkitykseen.

Mittausmenetelmien tulokset ovat yleisesti ottaen varsin kirjavia. fastText-algoritmin suorituskyky on yleisesti ottaen hieman parempi kuin perusmenetelmän; sen lisäksi sanaupotusten laatu on parempi. Leksikaalinen korvaustesti antaa parhaimmat tulokset, kun vain lähin naapuri otetaan huomioon. Sijamuodon todettiin olevan varsin tärkeä idiomin merkityksen määrittämiseen.

Mittauksien heikot tulokset voivat johtua monesta tekijästä, kuten siitä, että idiomien semanttisen läpinäkyvyyden aste voi vaihdella. Sanaupotusmenetelmä ei myöskään normaalisti ota huomioon sitä, että myös sanayhdistelmillä voi olla useita merkityksiä (kirjaimellinen ja idiomaattinen/kuvaannollinen). Suomen kielen rikas morfologia asettaa menetelmälle myös ylimääräisiä haasteita.

Tuloksena voidaan sanoa, että sanaupotusmenetelmä on jokseenkin hyödyllinen suomen kielen idiomien tutkimiseen. Testattujen mittausmenetelmien käyttökelpoisuus yksin käytettynä on rajallinen, mutta ne saattaisivat toimia paremmin osana laajempaa tutkimusmekanismia.

Avainsanat – Nyckelord – Keywords
word embeddings, idioms, idiomatic expressions, machine learning, multi-word expressions

Säilytyspaikka – Förvaringställe – Where deposited
Keskustakampuksen kirjasto

Muita tietoja – Övriga uppgifter – Additional information

Contents

1 Introduction

Multi-Word Expressions have received a lot of attention over the last few years. The definition of a Multi-Word Expression (MWE) varies a lot depending on the source; for some it is nearly synonymous to idiomatic expressions, others may include compound nouns and other complex words and possibly collocations. The heterogeneity of the concept may be a clue to the complexity of the topic. In any case, the processing of MWEs is has been considered a tricky problem[1] for some time.

A lot effort has been expended in the recent years and the topic of MWEs has been the subject of many special conferences and workshops (Hoang, Kim, and Kan, 2009; Anastasiou et al., 2009; Markantonatou et al., 2018; Parmentier and Waszczuk, 2019) and it has been a central topic in many others (Mitkov, 2017), just to mention a few.

1.1 Motivations

The starting point for the topic was the Finnish idiomatic expression *Ei ole kaikki inkkarit kanootissa.* A variation of this idiom is the following:[2]

(1) *Ei ole kaikki muumit laaksossa*
Not all of the Moomin trolls are in the valley

'He has a few screws loose'

'He has lost his marbles'

This connection lead to the realisation that idioms are not necessarily fixed and may exhibit considerable variation. In this particular case the phrase is called an *idiomatic construction.* This specific construction has been very popular[3] and has also been studied in detail in (Kortelainen, 2012) (see also chapter 2.2). The study of idioms eventually led me to the concept of *Multi-Word Expression.*

[1] Or, as titled in one of the early papers (Sag et al., 2002), "A Pain in the Neck for NLP".

[2] Slightly adapted Leipzig glossing rules (https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf) are used for listing the idioms in this thesis. The first line is for the idiomatic expression in Finnish, the second line for literal (English) translation and the third line (and possible subsequent lines) for a proper translation (i.e. free translation).

[3] https://www.kielikello.fi/-/kaikki-muumit-laaksossa

Another significant motivation for choosing the topic was to be able to work with/on Finnish language: First, despite efforts by many people working on other languages, English remains the first choice for any NLP work, including research on MWEs and idioms. Even today, this applies both to the academia - as noted in (Bender, 2019) - and the corporate world. Of the long papers in ACL conferences in 2016, 69 % did evaluation in English only.[4] In the non-academic setting, the language technology landscape (as part of the larger phenomenon of AI services) is largely dominated by US-based technology giants. Services are always provided first for English and then for other major languages. While Finnish cannot be considered as a so-called "low-resource language", support for Finnish in various services - when it exists - is at a lower level.

If using technology built on the assumptions of the English language, the rich morphology[5] and agglutinative nature of the Finnish language (Koskenniemi, Rehm, and Uszkoreit, 2012, p. 47) will produce additional challenges, especially if the methodology does not account for the internal structure of words. At the time when I started the thesis, there had been few, if any, computational studies of Finnish Multi-Word Expressions (to my knowledge). Therefore the combination of the complex Finnish morphology and MWEs was especially interesting as a topic.

Secondly, based on my personal experiences in Finnish technology world, there is considerable demand for language technology for Finnish. It was also important for me that the methodology to be used or developed during the thesis was cognitively motivated, that is, it should be based on solid evidence on how language is processed in the brain. Finally, it was my hope that I would be able to utilise methods based on neural networks.

1.2 Evolution of the Thesis

The focus of the study and the scope evolved over time. The original idea was to study the fossilisation of idioms,[6] with the assumption that (at least some) idioms start out as fully transparent and become more opaque over time

[4] https://sjmielke.com/acl-language-diversity.htm

[5] This is especially the case with nouns, with 15 grammatical cases. This is an important difference when compared to, for example, some major languages like Spanish and Portuguese where it is the verbs that have a large variety possible inflections (conjugations), not the nouns. See also appendix B and the URL http://jkorpela.fi/finnish-cases.html.

[6] The original title of the thesis was supposed to be "Fossils and Mumin Trolls".

(such as *kick the bucket*[7]). As news data has been found be a "good fit for studying language evolution" (Yao et al., 2018), I initially chose the Finnish newspaper corpus (1775-1917) from the National Archive as the primary data set for studying fossilisation.

This question of fossilisation was soon dropped, as was eventually the whole approach of studying idioms diachronically, leaving the data derived from public domain works in the Gutenberg project (see chapter 4) as the only data set. There were several reasons for this. First, it may take a long time for an idiom to change. Consider, for example, the idiom *Ei ole kaikki kotona* 'not everyone/everything is at home' (the precursor to *Ei ole kaikki muumit laaksossa*) whose origin likely predates written Finnish language (see chapter 2.2). The specific idiomatic structures studied - simple verbal phrase idioms - may be incredibly stable (as explained in chapter 2.8.1), which leads to a question whether studying semantic change for these types of idioms is a case of *Let's not go to Camelot, it is a silly place*,[8] that is, whether it makes much sense[9] - especially when the time periods in question are only in the order of decades.

Second issue (at least at the time) was the quality of the OCR'd data (Kettunen, Pääkkönen, and Koistinen, 2016). Thirdly were also concerns regarding the validity of the methods used to detect and then evaluate semantic change (see chapter 2.8).

Word embeddings were early on chosen to be the methodology to be used. The primary reasons for this were simplicity and computational efficiency - studies can be done without a large amount of computational resources, as is often the case with deep learning. The scope of this study was eventually narrowed down to pure synchronic analysis and two-word verbal idioms (V+N or N+V).

1.3 Research Questions

The general question since the choice of methodology was: how can idioms be studied with word embeddings, and how applicable embeddings in general are for this kind of study in the first place?

The first of the questions involve the measurements and tests that are to be

[7] https://en.wikipedia.org/wiki/Kick_the_bucket

[8] https://en.wikiquote.org/wiki/Monty_Python_and_the_Holy_Grail#It's_a_silly_place

[9] Note that this does not preclude studying other aspects, such as the productivity of idiomatic constructions, as has been done in (Petrova, 2011; Kortelainen, 2012).

developed, that is, compositionality and lexical substitution tests. How do they fare with the idioms found by Nenonen (see chapter 2.6). The question here specifically concerns the usefulness of these methods for determining the idiomaticity of the selected verbal idioms.

Another topic considers improvements to the standard word2vec embeddings. How do subword embeddings with fastText compare to the base case?

Other findings by Nenonen are also tested. How significant are various grammatical properties such as noun case for indexing meaning? Can "idiomatic minimal pairs" (where the noun case indexes idiomatic meaning) such as *vetää lärviin* / *vetää lärvit* (see chapter 2.6) be found?

Another important question is handling multiple senses. How should one properly distinguish between literal and idiomatic interpretations of Finnish MWEs?

1.4 Thesis Structure

The first chapter introduces the topic of the thesis and explains the motivations for doing the study. The research questions are also outlined.

The second chapter starts with the background for idioms and multi-word expressions, especially from the perspective of the Finnish language. The cognitive science background and its relation and relevance to the study of idioms is also touched upon. The chapter concludes with some background in distributed methods, including word embeddings.

The third chapter deals with methodology. Methods that have been used to investigate, classify and measure multi-word expressions are given an in-depth review.

The fourth chapter describes the Gutenberg data set and its limitations.

In the fifth chapter, the methodology of the thesis is developed, based on the related works and the background information regarding Finnish idioms. The first task or question of the thesis revolved around developing and evaluating methods for measuring idiomaticity, with the results being evaluated against a gold idiom list. The evaluation is done with both word2vec and fastText algorithms. After this classification is done with a neural network for a small labeled data set. Next section attempts to show the relevance of various grammatical properties for the meaning of the idiom. In the "odds and ends" part exploration with idiomatic minimal pairs and synonymy is done. The chapter ends with clustering for multiple senses.

In chapter six the results of the studies are evaluated against the research questions. The seventh chapter outlines conclusions drawn from the results.

Last but not least, the final chapter looks forward to whatever studies could be done as a result of this thesis. The various exclusions are evaluated and methods and ways to include them in future studies are also considered.

1.5 Acknowledgements

No man is an island and this thesis would not be possible without the help and guidance of many people. Most of all, I would express my gratitude to my supervisor and professor Jörg Tiedemann. I would also like to thank all the seminar participants for their constructive and helpful comments. Finally, many thanks to all those fellow students and teachers I had the pleasure of studying with at the University of Helsinki over the years.

2 Background

2.1 Definitions

Productivity The degree to which a grammatical process is used to coin (or derive) new expressions or forms. Productivity applies to all levels of language, including phonology, morphology, syntax and semantics (Säily, 2014, p.23). In this thesis, productivity, when applied to idioms, is largely used to refer to the extent of variation in idiomatic constructions and new idiom formation.

Compositionality The degree to which the meaning of an expression can be derived from the meanings of its components.

2.2 Into the Moominvalley

Kortelainen studied the idiom *ei ole kaikki muumit laaksossa* 'not all of the Mumin trolls are in the valley' in (2012), which is an example of the idiomatic construction *ei ole kaikki X:t Y:ssä* ('not all of X are in Y') and, based on a search of internet forums, traced the origin to 2005-2007, which makes the construction a relatively recent phenomenon. Kortelainen also ponders whether the equivalence of a variant of the construction - *ei ole kaikki inkkarit kanootissa* ('not all of the indians are in the canoe') - to the Swedish direct translation *inte alla indianerna i kanoten* is purely accidental or not.

Kortelainen found that this construction is very productive. It does have some limitations, though, as there needs to be a semantic connection between *X* and *Y*. Consider, for example, the phrase *ei ole kaikki hirvet zeppeliinissä*[10] ('not all of the elks are in the zeppelin'), which doesn't work because *elks* are not normally associated with *zeppelins*. The connection between the two varying components seems to be based on the semantic metaphor "MIND IS A CONTAINER" (Keysar and Bly, 1995, p. 91).

Finally, Kortelainen connects the construction to the idiom *ei ole kaikki kotona* 'not everyone/everything is at home', which was found in a book from 1644 (Kortelainen, 2012, p. 92-93) with its current meaning:

> Riginoldus: *Tosin ei sinulla ole caicki kotona.*

[10] Contributed by my colleague in a discussion regarding Finnish idioms.

This probably means that idiom actually predates written Finnish language - and has always been semantically transparent, unlike the English "poster child" for idioms, *kick the bucket*.

2.3 Multi-Word Expressions and Idiomatic Expressions

Various works have defined the concept of Multi-Word Expression in different ways, depending on the focus and needs of the study. Savary, Candito, et al. (2018) consider MWEs sequences of words, in which at least two components are lexicalised and some degree of idiosyncrasy is displayed. They classify verbal MWEs to three categories: Universal (Light Verb Constructions (LVC)[11] and idioms), Quasi-universal (inherently reflexive verbs and Verb-Particle Constructions (VPC)[12]) and other verbal MWEs.

Sag et al. (2002) categorise MWEs into lexicalised and institutionalised phrases, with the former divided into fixed expressions (such as *ad hominem*), semi-fixed expressions (which include non-decomposable idioms such as *kick the bucket*) and syntactically-flexible expressions. In (Siyanova-Chanturia, 2013), MWEs are loosely defined as "highly familiar phrases that exhibit a certain degree of fixedness".

Nenonen (2007) identifies four potential features for idioms. First, they consist of multiple words.[13] Secondly, they are non-compositional and thirdly (relatively) restricted regarding morphological, syntactic and/or lexical variation and finally they need to be conventionalised (or institutionalised[14]), which is a "social and psychological process" (Nenonen, 2002b, p.9). She further posits that idioms possess "conventional unexpectedness" (2002b, p.133), that is, they don't match the context or are otherwise used contrary to expectations.

Nunberg, Sag, and Wasow (1994, p.492-493) list six potential features for idioms: Conventionality, Inflexibility, Figuration (idioms often involve metaphors), Proverbiality (describing "a recurrent situation of particular social interest", Informality and Affect (idioms aren't usually used for neutral situations). Of

[11] The "lightness" coming from the notion that the verb (in this context) has little semantic content (Nenonen, 2007; Butt, 2010; Savary, Candito, et al., 2018).

[12] A combination of lexicalised head verb and its dependent particle (Savary, Ramisch, et al., 2017), such as "get up the hill".

[13] Even though it might be interesting to consider "single-word idioms", this restriction is nevertheless a linguistic convention. As the focus of the thesis in on Finnish, this established convention is followed.

[14] In linguistics, one definition of this is that the word is "accepted as a phraseological unit of the language" (Petrova, 2011, p.36)

these, only *conventionality* is considered obligatory.

Sheinfux et al. (2019, p. 41) argue that compositionality[15] is not "a primitive semantic property of idioms", proposing to classify idioms based on figuration and transparency. This is in line with earlier claims that the majority of idioms are actually semantically compositional (Keysar and Bly, 1995, p. 90; Nunberg, Sag, and Wasow, 1994, p. 491).

Semantic non-transparency has been identified as one of the potential properties of idioms. Transparency, usually defined by the closeness between the idiomatic and literal interpretations, can be considered to range from full transparency to opaqueness (Sporleder and Li, 2009; Conklin and Schmitt, 2012). Based on a study of Hebrew idioms, Sheinfux et al. (2019) found that transparency and figuration made idioms "more amenable it is to various transformations" (roughly correlating with the definition of "idiomatic productivity" used in this thesis). In (2009), Fazly, Cook, and Stevenson note that literal expressions are more likely to have lexical and/or syntactic variation.

Wulff notes in (2013, p. 285) that idiomaticity is an "inherently psychological construct". As the concept of *conventionality* has been identified as the only common criterion for idiomaticity (which, it deserves further elaboration. For idioms various definitions have been used: A linguistic regularity to which "a population has implicitly agreed to conform" (Nunberg, Sag, and Wasow, 1994); a phrase is "identified as idiomatic within a speech community"[16] (Hall, 2009); or more simply: "people know them" (Sidtis, 2009). At a risk of oversimplifying, in layman's terms conventionality refers to an expression that an (idealised) native / L1 speaker (in a Chomskyist "linguistic competence" fashion, perhaps) considers as idiomatic. As this is essentially a sociolinguistic phenomenon, it does not yield itself easily to computational approaches.

2.4 Cognitive Background

As *conventionality* related to how native speakers consider idioms, this provides a way to move from sociolinguistics to psycholinguistics. How do native speakers process idioms - especially compared to non-native speakers? This chapter reviews some cognitive studies on how, on one hand, regular words

[15] The term used by them is *decomposability*.

[16] The concept has been notoriously difficult to define within sociolinguistics. One definition of a speech community is "a group of people who speak in a distinct, identifiable style" (Milburn, 2015).

and, on the other hand, MWEs and idioms are processed in the brain.

Regarding regular morphological inflection, recent research from EEG, MEG and fMRI studies (Leminen, Jakonen, et al., 2016; Leminen, Smolka, et al., 2018) has affirmed that regular processing happens "online", i.e. word forms are decomposed before semantic processing is done, while irregular forms are handled directly according to the dual process model.[17] For derivational morphology the results are mixed and for compound words the studies are perhaps too scarce to draw any firm conclusions.[18]

2.4.1 MWE Processing in the Brain

Geeraert, Baayen, and Newman (2018) studied the impact of variation on MWE processing. The authors created four variants from the canonical form (example idiom *hear something through the grapevine*): lexical variation, partial form, integrated concept (where something is added to the idiom (*hear something through the judgemental grapevine*) and idiom blend (*get wind through the grapevine*). The various forms were tested for acceptability ratings and reaction times; eye-tracking methods were also used to gauge fixation times (i.e. how long the reader spent on processing various parts of the expression). The canonical form was considered the most acceptable form, while partial form was the least acceptable one. Regarding reaction times, the processing of lexical variation and idiom blends was not substantially slower. The length of the expression was found to be significant: longer idioms were more likely to be interpreted literally. The study also highlighted the relevance of predictability (and priming): idioms were faster to process in distinctive contexts. Based on the results, the authors questioned the dual-route model of language processing.

In another eye-tracking study (2017) on English V+N and V+particle MWEs, Yaneva et al. showed faster processing for formulaic phrases for both native and non-native speakers. Native speakers were found to have a processing advantage for the nouns in MWEs, which the authors attributed to higher exposure to English, theorising that the first word of the MWE was used for

[17] Dual process model posits that lexical and compositional access are done in parallel (form-with-meaning) (Geeraert, Baayen, and Newman, 2018; Leminen, Smolka, et al., 2018). Leminen contrasts this with the *two-stage* model, where decomposition comes first and semantic interpretation later (i.e. form-then meaning).

[18] This could be interpreted as being related to compositionality: (regular) morphological inflection is always compositional, while derivation is not necessarily so. For Finnish, the "forced" decomposition is also understandable as, for example, the number of possible forms a single noun can have is quite high and storing all of the forms wouldn't be feasible.

disambiguation.

Some have considered idioms (and by extension MWEs) to be a part of a larger phenomenon called *formulaic expressions*. According to various estimates between one third and one half of the expressions stored in long term memory are formulaic expressions (Conklin and Schmitt, 2012), thus making them very common. In line with the other research in this chapter, Conklin and Schmitt show that native speakers have an advantage when processing formulaic expressions.

2.5 MWEs and Senses

Most works considering the MWE senses assume that MWEs have (at most) two senses: literal and non-literal (Sporleder and Li, 2009), or literal and idiomatic (Katz and Giesbrecht, 2006; Cook, Fazly, and Stevenson, 2007; Fazly, Cook, and Stevenson, 2009; Conklin and Schmitt, 2012), although Cook, Fazly, and Stevenson acknowledge that both literal and idiomatic senses could have "multiple fine-grained senses". Nenonen (2007, p. 321) uses the labels literal and figurative (metaphorical, metonymic, idiomatic). While this classification is for a single word *silmä* 'eye', it's hard to argue why the same labels couldn't apply for MWEs as well. Petrova (2011, p. 96) notes that "idiomatic interpretation itself can vary depending on the context", while lamenting the lack of psycholinguistic studies on this subject.

When a comes to the proportion of literal vs non-literal expressions, Fazly, Cook, and Stevenson (2009) analysed 60 idioms and found that nearly half of the idioms could be interpreted literally. Of these, around 40 % of the examples were literal.

2.6 Idioms and Finnish Language

Based on various Finnish corpora, prototypical Finnish idioms consist of a verb phrase with a verb in a finite or infinite form and one or more complements (Nenonen, 2007). Particularly "idiom-prone" words are basic verbs and body part nouns. The most common verbs, also called nuclear verbs, are "pragmatically neutral" and can occur in many different contexts (Nenonen, 2007). These tendencies are also reflected in cross-linguistic studies (Niemi et al., 2013); idioms with body part nouns are also common in other languages (Nenonen, 2002a, p.114).

The most common verbs and nouns in idioms are shown in table 1. The 10

most common verbs listed account for half of the verbal idioms in the studied corpora.

Verb		Noun	
olla	'be'	silmä	'eye'
ottaa	'take'	mieli	'mind'
saada	'get'	pää	'head'
mennä	'go'	suu	'mouth'
pitää	'keep'	naama	'face'
vetää	'drag'	asia	'thing'
tulla	'come'	korva	'ear'
tehdä	'do'	aika	'time'
käydä	'fit'	sana	'word'
panna	'put'	turpa	'mouth'

Table 1: Most common verbs and nouns that participate in Finnish verbal idioms. Partially reproduced from (Nenonen, 2002b, p. 57)

Nenonen also points out that idiomatic usage is highly dependent on the inflection of the noun; partitive, illative and adessive cases are most common with nominative, partitive and illative following. When it comes to grammatical features, Nenonen finds that the plural form can be particularly idiom-prone, concluding that it may regarded "as an indexical marker of idiomaticity in Finnish" (when used to imply "unpredictable number"). Derivations and compounds are largely absent in these idioms, which Nenonen attributes to the apparent sufficient complexity of phrasal idioms.

The meaning of an expression may also depend on the case of the noun, as shown by the *idiomatic minimal pair*[19] examples 2 and 3 from (Nenonen, 2002a).[20]

(2) *vetää lärviin*
pull in the face.ILL
'to punch in the face'

(3) *vetää lärvit*
pull in the face.PL
'to get drunk'

[19] In phonology, a minimal pair is a pair of words that differ from each other by only one phoneme, yet they have different meanings. I've introduced the new term here to cover a similar phenomenon related to Finnish idioms.

[20] Examples originally from the Finnish comic *Viivi and Wagner*.

2.7 Idioms and Type of Language

There are considerable differences between different types of language and different registers. In (2010), S. Gries studied the British National Corpus Baby (which includes spoken and written data) and found that the spoken data in the corpus had "shorter sentences and more formulaic expressions". Given the demands of language production and understanding, this is to be expected (see chapter 2.4.1).

A major difference between spoken and written registers is that the latter can be planned, revised and edited. As a consequence, speech may be incomplete and more ungrammatical, while written text can be more complex and is more conformant to (grammatical) rules (Biber and Conrad, 2009, p. 85,109,117-118), in addition to having a "high type-to-token ratio" (essentially extent of vocabulary per length of text) and longer sentences (Louwerse et al., 2004). Speech is also interactive and more affective (Biber and Conrad, 2009, p. 85), making it more likely to contain idiomatic utterances.

The degree of formality in written material is also significant. More informal language can be expected in sources like internet forums, such as Usenet, which has been used as source material in many works, including (Petrova, 2011), social media and instant messaging. In her work on Finnish idioms, Nenonen (2007) chose to study juvenile books because they include "plenty of colloquial expressions and therefore idiomatic material". Idioms are especially prevalent in social media and internet blogs in a form that has been called *kirjoitettu puhekieli* (Kortelainen, 2012, p. 7) ("written speech"), i.e. written material that uses "conventions of speech" (Koskenniemi, Rehm, and Uszkoreit, 2012, p. 48).

In a crosslinguistic study of five European languages (Niemi et al., 2013) it was found that newspaper text is more non-literal than fiction. The authors reasoned that this is the result of a need to describe the world in fiction (i.e. world building), whereas in news the world is shared between the authors and readers, allowing the use of figurative language. Journalism is also known for "manipulating idiomatic expressions for humor or cleverness"(Fazly, Cook, and Stevenson, 2009).

2.8 Semantic Change

Various types of semantic change are listed in (Tahmasebi, Borin, and Jatowt, 2018, p.39). The usual suspects in this list include broadening and narrowing and adding new or related senses to words. Other types of changes include

borrowing, creation and loss (Bowern, 2019).

The study of semantic change with computation linguistics has picked up in recent years, especially with distributional methods like word embeddings (see next chapter). These methods have, among other things, affirmed the change in meaning for the word 'gay' (Hamilton, Leskovec, and Dan Jurafsky, 2016) and quantifying the appearance of word sense changes (Tahmasebi and Risse, 2017).

Various "laws of semantics" have been suggested based on these and other studies, such as law of innovation (polysemy is correlated with semantic change), law of conformity (frequency is correlated with semantic change) (Hamilton, Leskovec, and Dan Jurafsky, 2016) and law of parallel change (semantically close words experience similar changes) (Tahmasebi, Borin, and Jatowt, 2018, p.21).

Many of these findings have been criticised as potentially being based on data artefacts (Dubossarsky, Grossman, and Weinshall, 2017) and depending on various random factors and the order of the processing of data (Sommerauer and Fokkens, 2019).

2.8.1 Semantic Change of Idioms

While the semantic change of words has seen a lot of work, the literature of quantitatively investigating changes in idioms or MWEs is sparse. None of the surveyed studies on semantic change make more than a passing reference to idioms.

Butt (2010) explores light verb constructions (LVC) diachronically, the "lightness" coming from the fact that these verbs have little semantic content (in the specific context), which echoes the observation regarding semantically neutral verbs made by Nenonen in chapter 2.6. Despite the lightness, Butt finds that the verbs are "form identical", that is, they inflect normally just like the "main verbs".

Butt criticises the traditional view that the lightness came about by (gradual) diachronic change. Based on the Indo-Aryan language family (dating back 3000 years) and specifically languages like Old Bengali and Old Hindi (from around 1100 CE), Butt notes that LVCs "can be identified clearly and continually over thousands of years", concluding that these kind of constructions are "stable with respect to historical change".

2.9 Distributional Semantics

According to distributional semantics the meaning of a word is based on it's environment or distribution, or the set of contexts in which it appears (Daniel Jurafsky and Martin, 2018, p. 104). This can also be expressed with the often used quote by (Firth, 1957, p. 11): "You shall know a word by the company it keeps".

2.9.1 Word Embeddings

Vector-space models have been used for modelling word semantics for a while. The vectors in these models are commonly called embeddings, as the word is "embedded in a particular vector space" (Daniel Jurafsky and Martin, 2018, p. 105).[21] While these kinds of distributed representations had been used before, it was the word2vec model by Mikolov, Chen, et al. (2013; 2013) that popularised the approach. The model uses short and dense (real-valued) vectors. It has two variants: Continuous Bag-of-Words (CBOW) and Skip-Gram. Both approaches use supervised learning based on simple logistic regression: CBOW predicts the target word based on the context and Skip-Gram, conversely, the context words based on the target word.

The original model has been extended over the years to cover sentences (Salton, Ross, and Kelleher, 2016; Melamud, Goldberger, and Dagan, 2016) and documents (Le and Mikolov, 2014), among a few. Another model similar to word2vec is GloVe (Pennington, Socher, and Manning, 2014).

2.9.1.1 Effects of Hyperparameters

The size of the context window has an impact on the embeddings. Smaller windows tend to yield syntactic/functional similarities (that is, words occur in similar syntactic contexts) and bigger windows lead to more topical similarities (Goldberg, 2017, p. 128; Daniel Jurafsky and Martin, 2018, p. 118).

When it comes to the optimal context window size, in a word disambiguation study (Iacobacci, Pilehvar, and Navigli, 2016) the optimal window size was found to be 10. A study on the compositionality of German N+N compounds (Schulte im Walde, Müller, and Roller, 2013) found the optimal value to be 20. The general conclusion is that the - as with other hyperparameters - the optimal window size depends on the task.

[21] The term "embedding" apparently comes from linear algebra, but it's unclear who first used it in the current context.

Skip-Gram has generally been found to lead to superior results when compared to CBOW (Iacobacci, Pilehvar, and Navigli, 2016; Caselles-Dupré, Lesaint, and Royo-Letelier, 2018). Caselles-Dupré, Lesaint, and Royo-Letelier (2018) experiment with fine-tuning Skip-Gram hyperparameters, finding that it may make sense to deviate from the default values of negative sampling parameters.

2.9.1.2 Handling Morphology

One of the limitations of the base word2vec model is that the internal structure of the word is ignored. This limitation is particularly important for morphologically complex languages, including Finnish .

Subword embeddings (Bojanowski et al., 2017) are one of the most important and influential extensions in this regard to word2vec. In this method, each word is represented as a sum of the representation of the word's substrings (n-grams of length of 3-6 characters). The example given by Bojanowski et al. is the word *where*, whose 3-letter n-grams are *<wh*, *whe*, *her*, *ere* and *re>*, where *<* and *>* represent word boundaries. In addition to the n-grams, the word itself (*<where>*) is added to the vector.

In addition to improving the representation by sharing the subword vectors, subword embeddings have the advantage of being able to model out-of-vocabulary (OOV) words. Subword embeddings has been shown to have superior performance in many applications when compared to word2vec, although this varies by task and language (Salle and Villavicencio, 2018; Döbrössy et al., 2019; Zhu, Vulić, and Korhonen, 2019).

Another tool that is commonly used for modelling the internal structure of words is Morfessor[22] (Creutz and Lagus, 2005; Virpioja et al., 2013), which segments words probabilistically into morpheme-like units. Morfessor has worked especially well for morphologically complex agglunative languages like Finnish. The morpheme-like units produced by Morfessor have also been shown to correlate with brain activity in an MEG neuroimaging study (Hakala et al., 2018).

[22] https://github.com/aalto-speech/morfessor

3 Related Works

The methodology for dealing with MWEs can generally be dividing into two categories: 1) detecting when something is a Multi-Word Expression and 2) identifying the properties of the MWE (including sense disambiguation).

3.1 MWE Detection and Identification

3.1.1 Sequence tagging

Sequence tagging has commonly been used for Named Entity Recognition or NER, that is, extracting entities from text: names, places, times or dates and so on. An example from (Daniel Jurafsky and Martin, 2018, chapter 17):

> Citing high fuel prices, *[ORG United Airlines]* said *[TIME Friday]* it has increased fares by *[MONEY $6]* per round trip on flights to some cities also served by lower-cost carriers.

The standard form of sequence tagging uses BIO (or IOB) tagging, where B stands for the token in the beginning of the sequence, I for a token inside a sequence and O for a token outside of any sequence (ibid). Using this notation, a sample of the above quote would look like this:

> United (B-ORG) Airlines (I-ORG) said (O) Friday (B-TIME) it (O) ...

For tagging MWEs, the simplest approach has two deficiencies: 1) it cannot handle discontinuous expressions and 2) it doesn't account for the differences between weak and strong MWEs. Here *strong* is defined as clearly idiomatic and *weak* less so, e.g. mostly compositional collocations (Schneider, Danchik, et al., 2014). Three reasons are given for grouping non-contiguous tokens: 1) internal modifiers (***make*** *good* ***decisions***), 2) passive constructions (*they* ***gave*** *me* ***a break***) and 3) internal arguments (Schneider, Danchik, et al., 2014).

To account for these, additional tagging schemes have been created (ibid):

- 4 tags: B, $\tilde{\text{I}}$, $\bar{\text{I}}$ and O ($\tilde{\text{I}}$ and $\bar{\text{I}}$ for strong and weak expressions, respectively)

- 6 tags: B, I, O, b, i and o (lowercase variants for expressions within a gap)
- 8 tags: B, $\tilde{\text{I}}$, $\bar{\text{I}}$, O, b, $\tilde{\text{i}}$, $\bar{\text{i}}$ and o (combination of the above)

Sequence tagging has also been used for identifying MWEs in e.g. (Peters, Ammar, et al., 2017; Moreau et al., 2018).

3.1.2 Discontinuous MWEs

The effect of handling gappy/sparse/discontinuous MWEs on identification has been analysed in (Moreau et al., 2018). Three options were investigated: including expression words themselves in the context vector, counting multiple occurrences and context normalisation. The investigation produced mixed results. The authors noted that the options seemed to have opposite effects on the identification of continuous vs discontinuous MWEs, suggesting trade-offs. They finally lament the lack of a standard approach for computing a context vector for MWEs.

3.1.3 Other approaches

As for other approaches, Colson (2017) identified idioms from non-contiguous n-grams and built a search engine called "IdiomSearch". Hurwitz (2012) used a technique called "Text Isolation" (essentially morphological decomposition) to deal with Hebrew MWEs.

3.2 MWE Classification and Disambiguation

Peng, Feldman, and Vylomova (2014) divide idiom classification methods to two classes:

- type-based detection based on lexical properties (e.g. lexical and/or syntactic fixedness and non-compositionality)
- token-based detection (distinguishing between literal and non-literal instances)

Sporleder and Li (2009) identify idiomatic expressions based on lexical cohesion and topic mapping; an idiom is an "outlier" if it doesn't match the

surrounding topic. In this way, they consider the idioms similar to spelling errors, as in "semantic outliers that violate cohesive structure". The same idea was used in (Peng, Feldman, and Vylomova, 2014) to classify expressions with topic models and emotion intensity.

Birke and Sarkar (2006) use a "*nearly* unsupervised"[23] clustering algorithm called *TroFi* to distinguish between literal and non-literal uses of idioms. The algorithm is based on classifying whole sentences containing the target expressions.

Katz and Giesbrecht (2006) identify non-compositional MWEs by using Latent Semantic Analysis (LSA), distinguishing between non-compositional and compositional interpretations. The method is based on the expectation that compositional expressions occur in similar contexts to their components, whereas non-compositional expressions do not.

3.2.1 MWE Encoding

The topic of encoding MWEs within corpora was recently investigated in (Lichte et al., 2019). The authors note that the challenges that make MWEs difficult to process also extent to encoding them and advocate for the use of fully flexible encoding formats.

When it comes to encoding MWEs for downstream processing, the advise given in (Goldberg, 2017, p.133) to pre-process the text so that MWEs better fit "the desired definitions of words". The most common way to do this is by joining the component words with underscore ("_").

3.2.2 MWE Disambiguation with Word Embeddings

One of the key issues with word embeddings is that a single embedding is produced for each word form, thus conflating different senses. This also generally hurts the representation. Various methods have been tried to rectify this for sense disambiguation.

Unsupervised word sense discovery (WSD) is used in (Reisinger and J. Mooney, 2010) to cluster words based on their context vectors. In another paper that also predates word2vec, Huang et al. also cluster words based on their context vectors (window size 5) with spherical K-Means. The words are then relabeled in the corpus, after which the embeddings are retrained.

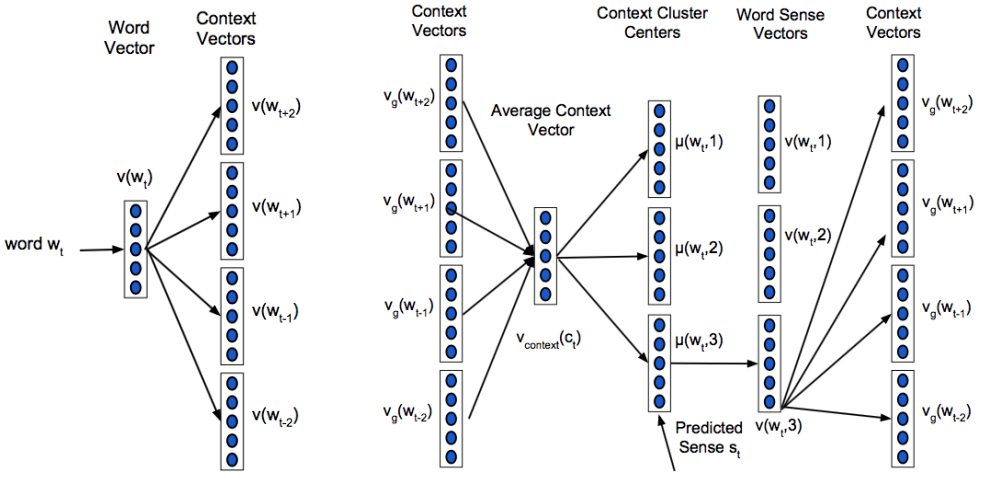
[23] The "nearly" in the title meaning using seed sets (literal and non-literal feedback), which makes it a "semi-supervised" approach.

Neelakantan et al. (2015) improve on this by jointly learning the senses while training the embeddings. Their model is referred to as Multiple-sense Skip-gram (MSSG) and the non-parametric variant as NP-MSSG (parametric meaning that the number of senses is determined in advance).

In this model, each word w has a global vector $v_g(w)$, with each sense of the word having an embedding (sense vector) $v_s(w, k)$ where $k \in 1, K$ (K is the number of clusters). The authors use global vectors instead of sense vectors to avoid computational complexity. Multiple embeddings are maintained per word type and the closest sense is selected during training (by finding the cluster center). A comparison of skip-gram calculation methods is shown in figure 1.

For the non-parametric variant, the clusters are learned during training. Initially there are no sense vectors or context clusters. After adding the first cluster for a word, a new cluster is created when the similarity between the observed context and any existing cluster is less than λ (a hyper-parameter) as shown in equation 1 ($\mu(w_t, k)$ refers to a cluster center).

$$s_t = \begin{cases} k(w_t) & \text{if } max_{k \in \{1, k(w_t)\}}\{sim(\mu(w_t, k), v_{context}(c_t))\} < \lambda \\ k_{max} & \text{otherwise} \end{cases} \tag{1}$$



(a) Standard Skip-gram (b) Multi-Sense Skip-gram (MSSG)

Figure 1: Comparison of models from (Neelakantan et al., 2015)

Building on (Neelakantan et al., 2015), Bartunov et al. build Adaptive Skip-gram (AdaGram) (2016), which uses Dirilecht process instead of clustering

to determine the number of prototypes.

Some of the more recent approaches have applied deep networks. Sentential context is used with bi-directional LSTM in (Melamud, Goldberger, and Dagan, 2016) to improve performance and provide disambiguation. ELMo (Embeddings from Language Models) from (Peters, Neumann, et al., 2018) is an LSTM model designed to be used with existing models. It also uses the whole sentence for the context.

The impact of various hyperparameters is investigated in (Iacobacci, Pilehvar, and Navigli, 2016). In their application - word sense disambiguation with IMS (It Makes Sense) system (Zhong and Ng, 2010) - the best performance is achieved with window size 5, number of dimensions 800 and exponential decay, where the importance of the context word is weighted according to the distance to the target word (in equation 2). Here W is the distance to the context word and the decay parameter α (in equation 3) is chosen so that the closest words contribute 10 times as much weight as the farthest words (window size $W = 10$).

$$w_{ij,x} = w_{ij}(1-\alpha)^{W-1} \tag{2}$$

$$\alpha = 1 - 0.1^{(W-1)^{-1}} \tag{3}$$

In (Schneider and Smith, 2015), MWEs are classified based on a set of WordNet supersenses (26 for nouns, 15 for verbs) such as PERSON, LOCATION, MOTION etc. Using the *Sent2Vec* algorithm, embeddings for sentences were created in (Salton, Ross, and Kelleher, 2016) to classify sentences as containing literal or idiomatic language.

3.2.3 Multilingual Approaches

Semantic mirror is based on the idea that different senses of a word map to different words in the target language (Dyvik, 2004). Moirón and Tiedemann (2006) used parallel aligned corpora (from Dutch to English, German and Spanish) to classify 200 MWE candidates as idiomatic or literal.

Hypothesising that compositional MWEs were more likely to have word-exact translation, Salehi, Cook, and Baldwin (2018) used PanLex (a massively multilingual corpus) to measure the degree of compositionality. Other parallel corpora that have been used for handling MWEs include OpenSubtitles2016 (Garcia, 2018) and the Bible (Tiedemann, 2018).

3.3 Measurements

Summarising from the introductory chapters, (measurable) MWE / idiom variation may be syntactic (word order may vary, expression may be non-contiguous), lexical (some components may be replaced with a word from the same class), morphological (inflection, conjugation, derivation), or semantic (continuums regarding transparency and non-compositionality) or a combination of these. This chapter outlines outlines some methods that have been used for measuring the degree of the variation. The measures fall roughly into two categories: fixedness and lexical association (including compositionality).

3.3.1 Measuring Fixedness / Inflexibility

Five generic tests were identified in (Savary, Ramisch, et al., 2017; Savary, Candito, et al., 2018) for measuring fixedness[24]: 1) cranberry words[25], 2) lexical inflexibility, 3) morphological inflexibility, 4) morphosyntactic inflexibility and 5) syntactic inflexibility.

Colson (2017) used *cpr score* (Corpus Proximity Ratio) to measure the syntactic variation of MWEs. The measure reflects the average distance between the components of an n-gram. Examples include *at the drop of a hat*, where the score is 1.0 (no gaps exist in the corpus in the middle of the MWE) and *Add insult to injury* with a score of 0.96 (i.e. some gaps exist for this expression in the studied corpora).

Fazly, Cook, and Stevenson (2009) developed lexical and syntactic fixedness measures for verb+noun idiomatic combinations (VNIC) based on a modified version of Point-wise Mutual Information (see equation 5) and Kullback-Leibler divergence. A unified fixedness measure is shown in equation 4, where F_{lex} and F_{syn} represent lexical and syntactic fixedness measures, respectively and α is a weighting factor.

$$F_{overall}(v,n) = \alpha F_{syn}(v,n) + (1-\alpha)F_{lex}(v,n) \tag{4}$$

[24] The papers actually consider these as testing non-compositionality, but I felt that they fitted more properly under the "fixedness" umbrella; therefore they are listed here.

[25] In this instance, we are essentially talking about fossil words, or words that don't appear outside the specific expression (Nenonen, 2002b, p. 123). The term "cranberry" was originally used in relation with morphemes. The origin of the term comes from the fact that the morpheme *cran* only appears in the word *cranberry*, making it a fossilised morpheme, see http://www2.let.uu.nl/Uil-OTS/Lexicon/zoek.pl?lemma=Cranberry+morpheme.

3.3.2 Measuring Lexical Association and Semantic Relatedness

Hoang, Kim, and Kan (2009) compared a great number of lexical association scores for modeling the degree of association between components. They divided them roughly into two classes: institutionalisation (whether a phrase is part of a semantic unit) and non-compositionality.

The first class includes traditional measures such as *Point-wise Mutual Information* or *PMI*. As this measure is correlates with frequency, favoring phrases where constituents have low frequencies, Hoang, Kim, and Kan suggest normalisation with penalisation terms based on marginal frequencies, e.g PMI divided by $NF(\alpha)$ or $NFmax$ in equations 6 and 7. In these bigram equations, N is the total number of bigrams and $f(X)$ refers to the frequency (token count) of bigram X.

$$\begin{aligned} PMI(x,y) &= log\frac{P(xy)}{P(x*)P(*y)} \\ &\approx log\frac{Nf(xy)}{f(x*)f(*y)} \\ &\quad \text{where } * \text{ refers to all words} \end{aligned} \tag{5}$$

$$\begin{aligned} NF(\alpha) &= \alpha P(x*) + (1-\alpha)P(*y) \\ &\quad \text{with } \alpha \in [0,1] \end{aligned} \tag{6}$$

$$NFmax = max(P(x*), P(*y)) \tag{7}$$

The authors note that most context-based measures do not fare that well with detecting VPCs and LVCs (see chapter 2.3 for definitions), owing this to the high frequencies of particles in these expressions.

Gries and Wahl (2009; 2018) explore a concept called "Lexical Gravity" developed by Daudaravičius (2004). This measure weights collocational probabilities based on type frequency.

Wulff (2013, p. 281) used collocations to calculate "R" score for determining compositionality of an expression.

$$R_1 = \frac{n(W,C)}{n(C)} \tag{8}$$

$$R_2 = \frac{n(W,C)}{n(W)} \tag{9}$$

$$R = R_1 + R_2 \tag{10}$$

In equations 8 and 9 $n(W, C)$, $n(W)$ and $n(C)$ represent the number of collocates shared between word W and construction C, number of collocates for word W and finally for construction C, respectively. Equation 8 measures how much "the semantics of the construction is accounted for by the component word"; equation 9 reflects how much "of itself each component word brings into the constructional meaning". The final measure in equation 10 is a combination of these.

Sporleder and Li (2009) use *Normalised Google Distance* (Cilibrasi and Vitányi, 2007) to measure semantic relatedness based on the Google's page counts, e.g. comparing counts returned by "fire" and "coal" to those returned by "fire AND coal".

In (2015), Salehi, Cook, and Baldwin apply the Multi-Sense skipgram from (Neelakantan et al., 2015) to MWEs. They developed two measures for compositionality:

$$comp_1(\mathbf{MWE}) = \alpha sim(\mathbf{MWE}, \boldsymbol{C}_1) + (1 - \alpha) sim(\mathbf{MWE}, \boldsymbol{C}_2) \quad (11)$$

$$comp_2(\mathbf{MWE}) = sim(\mathbf{MWE}, \boldsymbol{C}_1 + \boldsymbol{C}_2) \quad (12)$$

where **MWE** is the vector associated with the MWE, $\boldsymbol{C}_i$ is the vector associated with the ith component word of the MWE, *sim* is a vector similarity function (such as *cosine*), and $\alpha \in [0, 1]$ is a weight parameter. To account for the variation in word forms, the maximum compositionality value is used for the whole MWE.

Character-level neural networks (LSTM) are used to measure compositionality in (Parizi and Cook, 2018) on three English and German datasets. The equations used for the measurements are 11 and 12. Additionally the compositionality of single component words is calculated with equation 13.

$$comp(\mathbf{C}) = sim(\mathbf{MWE}, \boldsymbol{C}) \quad (13)$$

The authors note that the kinds of models they used do capture some aspect of compositionality with the added benefits of predicting compositionality for OOV and low frequency expressions. They hypothesise that methods based on character-level neural networks may be complementary to other methods.

Noun-noun compound compositionality was measured in (Dhar, Pagel, and Plas, 2019). Several measures were evaluated: Similarity between compound constituents, similarity of the compound with its head or modifier, log-likelihood ratio (LLR), PPMI and Local Mutual Information (LMI). LLR

and LMI were found to be the best predictors, while similarity between constituents and similarity of the compound with the head also fared well.

In (Vecchi et al., 2016) additive, multiplicative, dilation methods and lexical function (i.e. matrix calculation) are used to obtain a vector representation for adjective+noun combinations. These representations are used to predict the acceptability / plausibility of various combinations for human judges. While the authors find similar quantitative performance for all models, they conclude that the lexical function is most appropriate for this task based on qualitative evidence.

4 Data Sets

There are several corpora that have been popular for studying Finnish idioms. Usenet, i.e. "internet news" has been used for many studies (Petrova, 2011; Kortelainen, 2012). The corpus built from the suomi24 discussion forum (2014) has also been a common choice. As the initial focus of this thesis was on semantic change of idioms, I became interested in the Longitudinal Corpus of Finnish Spoken in Helsinki (2014), which contains transcribed interviews from three decades. I settled on the digitised archive of Finnish newspapers from 1771 to 1917 (Kettunen, Pääkkönen, and Koistinen, 2016).[26]

To develop the basic methodology I chose to use the Finnish language books from Project Gutenberg[27], which hosts books in the public domain for various languages. As the diachronic aspect was eventually dropped from the scope, in the end the Gutenberg data set was also used for the whole thesis - mainly out of convenience, not because it is a particularly good data set. The fact that it had not been used as a data set before (perhaps for a good reason) was also interesting.

The data set contains both fiction and non-fiction; Finnish and translated works and books from various time periods. As the data is in public domain, it is also relatively old. All in all, it is a very heterogeneous data set.

4.1 Challenges and Limitations

The lack of annotated training material for Finnish MWEs for the most part precluded using supervised learning for handling MWEs. This influenced and limited the choices for the methodology. A small labeled set for supervised classification was created in this study, though - see chapter 5.2.5.

For studying idioms, a more informal / colloquial data set might have been more useful, as such data sets are more likely to contain idiomatic expressions in their canonical form (Nenonen, 2007, p. 312). However, the data set also contains works of fiction, which are difficult to classify when it comes to registers (Biber and Conrad, 2009, p. 132). These texts commonly include conversation pieces and thus potentially more idiomatic material. The prevalence of idioms in texts that were originally written in Finnish versus

[26] http://www.dlib.org/dlib/july16/paakkonen/07paakkonen.html

[27] The Gutenberg data for Finnish downloaded according to the instructions in http://www.gutenberg.org/wiki/Gutenberg:Information_About_Robot_Access_to_our_Pages in February 2019.

translated is unknown.

A more important limitation is the age of the material: the idiom list used for evaluation is modern / contemporary language while the language in the Gutenberg data set is generally much older.

Furthermore, as part of the material is fiction, it may contain informal (spoken) or dialectal forms or slang which are not captured by the methodology. Examples of these are *tehkäät*, *tehtäis*, *tehtihin*, *tehdäkkään*; *vetäsi*, *vetäis*; *ottakaat*, *ottais*, *ottakaas*.

4.2 Focus of Study

For simplicity and to limit the amount of idioms to analyse, the focus was narrowed to two-word verbal idioms (V+N or N+V). Non-contiguous phrases, that is, idioms that have one or more words in between (like *pitää tarkasti silmällä* 'keep close eye on') were also excluded, again primarily to constrain the amount of effort. As the task in this thesis is not about MWE identification or accounting for the full range of variation, I deemed narrowing the scope reasonable.

The exclusions were also also somewhat motivated by the cognitive background on MWE processing (see chapter 2.4.1) for various reasons: Contiguous "canonical" forms are expected to be more "formulaic". There are also limitations to how much variation is allowed before an expression is no longer detected (by a native speaker) as idiomatic. Non-canonical variations of the idioms can also be expected to have a much lower frequency.

The idioms to study were chosen based on a 'gold' list of idioms from (Nenonen, 2002b, p. 149-182). This list of 3354 idioms was narrowed to 2163 two-word idioms and further to 1259 verbal idioms (V+N or N+V). The list is available in github.[28]

[28] https://github.com/dustedmtl/thesisdata

5 Methodology

5.1 Preprocessing

The Gutenberg data had three different encodings: UTF-8, Latin1 and "ASCII" (i.e. scandinavian characters were encoded with two characters, e.g. *ä* -> *ae*). The seven files with the last encoding were discarded, as there is no certain way (absent proper morphosyntactic analysis) to convert all of the words correctly; for example *haen* and it's "possibly corrected" form *hän* are both valid word forms in Finnish. The Latin1 files were converted to UTF-8. A number of duplicate files were also removed (both UTF-8 and Latin1 versions existed). The total number of files (books) was 1963.

The data was lowercased and all punctuation removed. The sentences were tokenised using NLTK[29]. Some metadata was also captured for Gutenberg data (mainly the author). After this processing there were (roughly) 6 million sentences with 65 million words and a vocabulary size of 2 million.

The only character correction for the data was the *w* -> *v* conversion recommended in (Kettunen, Pääkkönen, and Koistinen, 2016).[30] The process of fixing is details in appendix A. There were 19677 fixed word forms in total.

No syntactic analysis was performed on the data. The primary reason for this was that I expected to use noisy OCR'd newspaper data from the 19th century as the main data source.

For analysis, only bigrams with the combinations V+N and N+V are taken into account. The voikko[31] library is used to get the list of possible analyses for each word form. The number of analyses per class (i.e. how many word forms can be analysed as member of the class) is shown in figure 2, while the statistics for the number of distinct analyses a word form can have is shown in figure 3. The figures include forms that may be analysed multiple times (for example, as a noun, *silmääni* may have partitive or illative case). The number of unique word forms is around 176000 for verbs and 954000 for nouns.

[29] http://www.nltk.org

[30] A major orthographic difference between modern Finnish and e.g. 19th century Finnish is the use of letter *w* instead of *u*.

[31] https://voikko.puimula.org. The morphological analysis is essentially done with omorfi: https://github.com/flammie/omorfi

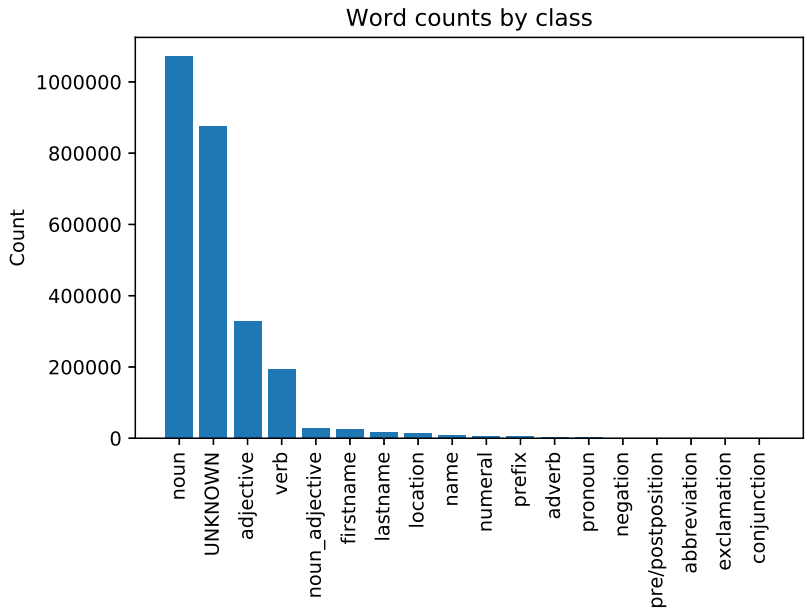


Figure 2: Number of analyses per class for the Gutenberg data

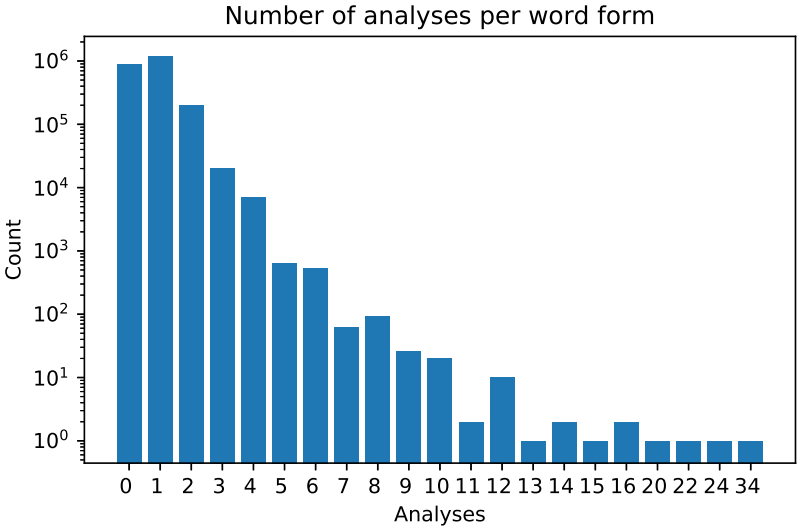


Figure 3: Number of analyses per word form for the Gutenberg data

Table 2 shows the most common unknown words: in the left side for all words, and in the right side for words that start with 'silm' (matches are expected to be variants of 'silmä'). The words roughly fall into three categories: 1) words that nowadays written separately - *niinkuin* (*niin kuin*), 2) spoken forms - *mun* (modern Finnish *minun* 'mine', i.e. 'belongs to me') and 3) archaic forms - *kauvan* (*kauan*), *silmäinsä* (*silmiensä*), *silmihin* (*silmiin*). For the last two inflections of *silmä*, the counts for the modern forms were 2875 and 15994, respectively.

All		silm*	
Form	Count	Form	Count
niinkuin	146569	silminnähtävästi	2088
ikäänkuin	76056	silmäinsä	1699
ennenkuin	72344	silmäs	699
sitte	63484	silmäini	508
sentähden	40409	silmäimme	288
mun	25097	silminnähtävä	247
mut	22740	silmäns	214
kauvan	17906	silminnähtävää	208
sun	17145	silmihin	204
nämät	14968	silmälläpitäen	159

Table 2: Words unknown to voikko: all words and words that start with 'silm'

Having so many unknown forms decreases the quality of the embeddings, as the size of the vocabulary is increased. If these unknown forms were properly accounted for, then words like *niinkuin* and *mun* would be handled as stop words and omitted from the data. The variants of the noun *silmä* would be relevant to the extent that they would be included in the bigrams to be analysed.

To be included in the list of bigrams for analysis, it was necessary that the word form could be interpreted as a verb or a noun, not that was necessarily tagged as a verb or a noun in the sentence (since no syntactic analysis is used). For example, *tulen* can be interpreted as a noun and as a verb and the particle *juuri* as a verb.

(4) *tulen*
fire.SG.GEN
I come.1SG.PRES

'Of fire'

'I come'

(5) *hän juuri*
he uproot.3SG.IMP

Likewise, the phrase *voi voi*[32] is extremely likely to be an interjection; nevertheless, it could be interpreted as a V+N or N+V combination:

(6) *voi*
butter
he can.3SG.PRES

In another example the phrase

(7) *tikku silmään*
a stick to the eye

based on the phrase

(8) *tikulla silmään sitä joka vanhoja muistelee*
a stick to the eye to whom reminisces the old

'let bygones be bygones'

is erroneously[33] included as N+V as *silmään* can also be interpreted as a verb (lemma *silmätä*). There were 30551 word forms in the data that could be analysed both ways (as a verb or a noun). The impact of these homonyms on the results is uncertain. The removal of punctuation also increased the possibility that some instances of the verbal idioms were misidentified.

Additionally, word forms that were included in NLTK's list of stop words for Finnish, Swedish or English were excluded. The use of Finnish stop words also had the effect of removing any forms of the Finnish verb *olla* 'to be', which is the most prolific verb for Finnish verbal idioms.

[32] https://en.wiktionary.org/wiki/voi#Finnish
[33] Although it is also an idiomatic phrase, it is not a *verbal* idiom.

5.2 Embeddings

5.2.1 MWE Encoding

The components of bigrams to be analysed are joined with underscore ("_"). To be able to train both the bigram and its components, the sentence containing the bigram is trained twice: first with the original sentence and then with the sentence fragment with the bigram including its context window.[34]

For example, for the sentence

> *Poika oli astunut taaksepäin ja tuijotti häneen jääkylmin katsein; ja vaikka Olina oli luonut pistävät silmänsä häneen, käänsi hän silmänsä pois pelvolla.*

in *Uusia kertomuksia* by *Magdalena Thoresen* in the Gutenberg corpus the bigram *pistävät_silmänsä* is trained using the fragment *ja vaikka olina oli luonut pistävät_silmänsä häneen käänsi hän silmänsä pois*. This inaccuracy would not exist if the neural network was built from the ground up e.g. on top of pytorch with a proper training for the bigram (only). This likely has the effect of making the expressions seem more compositional than they should be, as the context words appear more frequently with the bigram.

5.2.2 Training

The word2vec implementation of the gensim[35] library is used for training the word embeddings. Skip-gram with negative sub-sampling and hierarchical softmax (Mikolov, Sutskever, et al., 2013) is used with context window size 5, minimum word count 3, number of dimensions 100 and 5 training epochs. For most hyperparameters, the choice was the use defaults unless there was a compelling reason to choose something else (e.g. based on chapter 2.9.1). The chosen window size was expected to be more conductive to syntactic similarities, which is likely what we would be looking for in this task.

[34] One of the original word2vec papers (Mikolov, Sutskever, et al., 2013) includes a mechanism for detecting and training phrases. This method, however, does not appear to include the training of the original sentence with the components of the bigram as single words.

[35] https://radimrehurek.com/gensim/

5.2.3 Measurements

Of the various properties listed in chapter 2.3 that can be used to describe idioms, non-compositionality and lexical inflexibility can be measured (syntactic fixedness testing is omitted due to the preprocessing methodology). Most in-depth analysis is done for idioms that include the noun *silmä* 'eye', as it is the most idiom-prone noun in Finnish (Nenonen, 2007). Each idiom to analyse is measured with two methods. First, the compositionality score is calculated based on equation 12 on page 27, reproduced here for clarity:

$$comp(\mathbf{MWE}) = sim(\mathbf{MWE}, \boldsymbol{C}_1 + \boldsymbol{C}_2) \tag{14}$$

Secondly, a lexical substitution test is done. A number of nearest neighbours of the idiom are fetched using word2vec's *most_similar* function. If a word form is found in this list where one of the components is kept unchanged and the lemma of the other component is not the same as in the idiom, lexical substitution is considered to be possible and the form is considered less idiomatic. The tests are run for a variety of neighbourhood sizes. Table 7 in chapter 6.4 shows some examples of these (neighbourhood size 100).

5.2.4 Evaluation

5.2.4.1 Idiomaticity

The idiomaticity methods were evaluated against a 'gold' list of idioms (see chapter 4.2). For detailed analysis, the idioms were limited to those where the noun is *silmä* ('eye'). Additionally, only V+N forms were considered as N+V forms are much rarer in Finnish (Nenonen, 2002b, p. 29). Idiomaticity can also expected to be sensitive to word order - variants where the order of the words is changed should be less recognisable (again, see chapter 2.4.1 on MWE processing).

True positives and false negatives were evaluated based on a list that only included forms that actually occurred in the data with sufficient frequency (minimum count 5, by default). These idioms and their English translations are listed in table 3.[36] Figurative and metaphoric translations are listed as *figurative*. The compositionality and lexical substitution evaluations were done both on the basis of word form and noun case.

[36] I wasn't familiar with some of these idioms, so finding the non-literal proper translations took some effort. This is evidence for the notion that no speaker knows all of the idioms of their native language.

Idiom	Literal translation	Non-literal translation
ahmia_silmillään	to devour with one's eyes	figurative
avata_silmänsä	to open one's eyes	figurative
iskeä_silmää	to punch [other's] eye	to wink
kääntää_silmät	to turn one's eyes	figurative
miellyttää_silmää	to please one's eye	to find pleasing
pestä_silmänsä	to wash one's eyes	figurative
pistää_silmiin	to stick in the eyes	to stick out
pistää_silmään	to stick in the eye	to stick out
pitää_silmällä	to keep on one's eye	to keep an eye on
ristiä_silmänsä	to cross one's eyes	the make a cross at the eye level
saada_silmiinsä	to get in one's eyes	figurative
sattua_silmään	to hurt the eye	to stick out
ummistaa_silmänsä	to close one's eyes	figurative
uskoa_silmiään	to believe one's eyes	figurative
viehättää_silmää	to allure the eye	to find alluring

Table 3: Idioms to analyse with English translations

For both methods, false positives were evaluated by testing all applicable V+N forms against the gold list. While this list may be the most authoritative, it is not complete, thus some expressions that are marked as false positives would be considered idiomatic by a native speakers.

Finally, F1 (harmonic average) scores are calculated based on the recall and precision values.

5.2.4.2 Subword Embeddings

As noted in chapter 2.9.1.2, subword embeddings are particularly useful for two reasons. First, they are able to deal words not in the vocabulary (OOV). Secondly, it improves the representations and generally improves the performance.

Here the task is not about predicting unseen idioms, so the first reason does not apply. As for the second, the methodology might also make different noun forms closer than they would otherwise be (which is what we would not want, if noun case is relevant for meaning).

To check whether subword embeddings improve the results or not, compositionality and lexical substitution tests were also run on a model built on

gensim's implementation of fastText[37]. The model parameters were otherwise the same as with standard word2vec.

5.2.5 Classification

The classification set was based on sentences where the selected idioms with *silmä* were present. There were 21779 such sentences. This set was narrowed down to 4878 sentences containing idioms from table 3 and finally to 3807 ones where the noun case was the correct one for the idiom. These sentences were classified as either literal or non-literal (idiomatic, figurative, metaphoric) by the author (with native linguistic competence).

The classification was done with a neural network built on pytorch[38] with two hidden layers with dimensions 200 and 50 and one dropout layer with $p = 0.1$, 10-20 training epochs, learning rate 0.005 with stochastic gradient descent (SGD). Since the data set was fairly small, the training used 10-fold cross-validation and the results were averaged over 5 runs. Evaluation was done with context window sizes 5 and 10 and using both average and exponential decay (see equations 15 to 17) for calculating the context (that is, four sets in total).

The calculation of the context is based on the embeddings of the surrounding words. By default, the context is taken from the average of the contexts surrounding the expression:

$$\boldsymbol{v}_m = \frac{1}{2k}\Big(\sum_{\substack{j=m-k \\ j\neq m}}^{m+k} \boldsymbol{w}_j\Big) \tag{15}$$

where $\boldsymbol{w}$ is the vector representing the sentence, $\boldsymbol{w}_i \in \mathbb{R}^d$ is the embedding vector in position i, $\boldsymbol{w}_m$ is target MWE and k is window size. If the window were to extend beyond the boundaries of the sentence, the equation is modified accordingly, i.e. if the target MWE is the third token in the sentence, only the first two words are included in the context calculation (on that side of the word). If there are no context words (i.e. the sentence only contains the MWE), the sentence is omitted from further analysis.

[37] https://github.com/facebookresearch/fastText
[38] https://pytorch.org

Adapting from equations 2 and 3 on page 24, the context with exponential decay is calculated with equations 16 and 17 (with the equations modified based on the length of the available context, if necessary). The result is essentially a weighted average of the embeddings of the surrounding words.

$$\boldsymbol{v}_m = \sum_{\substack{j=m-k \\ j \neq m}}^{m+k} \boldsymbol{w}_j (1-\alpha)^{|m-j|-1} \tag{16}$$

$$\alpha = 1 - 0.1^{(k-1)^{-1}} \tag{17}$$

5.3 Exploratory Testing

The importance of noun case and other grammatical properties is tested for a number of idioms. The test is based word2vec's similarity function for average in-group vs out-group similarity, i.e. are members of a group (e.g. bigrams with a certain form / case etc) closer to each other than members of a different group. This analysis was done only with the base word2vec algorithm.

In a more exploratory vein, nearest neighbours for a number of idioms are listed. The list includes the idiom *tehdä mieli*[39] is compared to a variant *mieli tekee*; see examples 9 and 10.

(9) *tehdä* *mieli*
to make a mind
'to feel like [doing something]'

(10) *mieli* *tekee*
the mind makes
'to feel like'

An attempt is also made to find idiomatic minimal pairs similar to those found in examples 2 and 3 in chapter 2.6.

[39] There is also the phrase *mieleni minun tekevi* from the Finnish national epic *Kalevala*, which unfortunately cannot be tested as *tekevi* is not recognized by voikko.

5.4 Clustering

Clustering has been used in prior for disambiguation (see chapter 3.2.2), that is, distinguishing between literal and non-literal interpretations.

While a small labeled gold set was created in this thesis, it would be useful to see how well clustering works for disambiguation. The hoped-for result is that the instances of the MWE are correctly classified into literal and non-literal forms and the range of compositionality scores should increase. In concrete terms, the lowest compositionality score of the "idiomatic" cluster should be higher than the highest compositionality score of the "literal" cluster.

A key question here is: what are we clustering? In prior works the components themselves are clustered, but given the focus and methodology of this thesis this could lead to an explosion of forms (at least without lemmatisation) - especially since the words (both verbs and nouns) that generally participate in idioms are generally very polysemous. The rich morphology of the Finnish language would not help here.

The choice, then, is to cluster the bigram itself into two clusters. For simplicity, the approach from (Huang et al., 2012) is used, that is, to cluster, relabel and retrain. The specific focus is on idioms that, based on labelling (see chapter 6.6), seem to have a fair balance of literal and non-literal interpretations. Only one idiom - *ummistaa silmänsä* - seemed applicable, so it was chosen as the target expression.

The procedure is as follows: the instances of the idiom (based on nominative case) are clustered based on the embeddings of the surrounding context words - the same approach as used in chapter 5.2.5. Exponential decay (see equations 16 and 17 on the preceding page) is used for the calculation of the embedding/context, with fastText used as the embedding model. The instances are then clustered with KMeans. The corpus is relabelled based on the clustering, that is, instances of the idiom *ummistaa_silmänsä* are replaced with the clustered instances (*ummistaa_silmänsä:1* and *ummistaa_silmänsä:2*). The corpus is then retrained according to methodology in chapter 5.2.2, after which compositionality scores can be calculated separately for each cluster.

6 Results

6.1 Compositionality Scores

Most and least compositional bigrams (collocations) are shown in table 4, with lower score = more compositional. The minimum value is 0, which means full compositionality (the maximum value is above 1 due to the methodology). There appears to be some correlation between low frequency and the scores on both the high end and low end exhibit; in table 5 an equivalent result is shown for bigrams with a minimum count of 10. In both tables the most compositional expressions seem to be frequent collocations.

Similar analysis is shown for verbal idioms in table 6 where the second word is a form of the lemma *silmä* 'eye'. Here the effect of frequency is only exhibited for bigrams with low compositionality.

Most compositional			Least compositional		
Form	Score	Count	Form	Score	Count
maanpitäjä_pillastuko	0.0281	4	hovipoika_menee	1.2288	4
ärjytähän_vihapäässä	0.0292	3	kysyi_suutarinemäntä	1.2145	4
nahkaruoskalla_napauta	0.0339	5	kerttu_naurahtaen	1.1803	3
mesiheinin_herkuttele	0.0379	3	yössä_valvon	1.1638	5
suorimasta_surmiansa	0.0391	3	muutu_muotoon	1.1633	3
pohjani_porotan	0.0424	5	harhama_värähti	1.1557	3
helkyttele_hietarinta	0.0435	5	prinssi_auttakaa	1.1553	3
valjastele_varsojasi	0.0465	4	kantapoika_menee	1.1494	4
lihoilla_väiky	0.0481	4	tahdon_sir	1.1446	3
venymästä_vehnäsille	0.0487	3	morsian_lähtiessänsä	1.1318	4

Table 4: Most and least compositional bigrams.

Most compositional			Least compositional		
Form	Score	Count	Form	Score	Count
raitista_ilmaa	0.0496	493	narri_laulaa	0.9415	14
nurmilintu_väsy	0.0545	11	herra_suojaa	0.7015	13
täyttä_laukkaa	0.0553	746	koossa_pysymään	0.6933	11
silmät_tuijottivat	0.0731	186	lääkäri_vastoin	0.6885	25
hiki_valui	0.0733	165	katsokaa_peiliin	0.6835	11
nuku_nurmilintu	0.0764	12	ajat_takaa	0.6771	13
iätä_iästä	0.0789	15	pistää_silmiin	0.6766	13
täyttä_nelistä	0.0790	70	kohautti_olkapäitään	0.0792	619
tekee_kunniaa	0.6766	16	huomaa_kaikesta	0.6664	11
kostu_korpi	0.0794	31	voi_kauhistusta	0.6663	19

Table 5: Most and least compositional bigrams, minimum count 10

Most compositional			Least compositional		
Form	Score	Count	Form	Score	Count
silmät_tuijottivat	0.073	186	silmän_päästään	0.751	4
silmät_kiiluivat	0.081	116	tikku_silmään	0.678	3
hehkuivat_silmät	0.083	30	pistää_silmiin	0.677	13
silmät_säihkyivät	0.085	140	silmissä_välähtelee	0.671	3
verestävät_silmät	0.093	13	pistä_silmään	0.667	8

Table 6: Most and least compositional bigrams for V+N / N+V, N=*silmä* 'eye'

6.2 Lexical Substitution

Table 7 shows how the lexical substitution test works. The idiom *pistää_ silmään* is classified as idiomatic since most similar bigrams do not include ones with form *X_ silmään*. Here the range of the score is $[0, 1]$, where the score of 1 means that the forms are identical in meaning.

Form	Lexical sub	Same lemma
pitää_silmällä 'to keep an eye on'	tarkasti_silmällä (0.7689) heitä_silmällä (0.7967) salaa_silmillä (0.6725)	piti_silmällä (0.7231) pitäen_silmällä (0.7115) pitämässä_silmällä (0.7033)
avasi_silmänsä 'opened their eyes'	aukaisi_silmänsä (0.9045) nosti_silmänsä (0.8362) 'raised their eyes' painoi_silmänsä (0.7657) 'lowered their eyes'	
pisti_silmään 'stuck [something] in the eye'		pistivät_silmään (0.0520) pisti_silmiin (0.6240)

Table 7: Neighbours for various "silmä" idioms for lexical substitution test, neighbourhood size 100

6.3 Evaluation with Compositionality Scores

The classifications for idiomatic forms for noun *silmä* are shown in table 8. In this table and all subsequent ones, **Pos** refers to true positives and **Neg** to false negatives. The compositionality score cutoff used in this and subsequent tables is 0.42, as it provides the best performance (see figure 4 on page 46). Above this value, a form is marked as idiomatic. The minimum token frequency for an idiom form is 5 - if, for example, the total column has the value 3, then there are 3 forms that occur at least 5 times. For *pistää silmiin* these forms are *pistää_ silmiin*, *pistivät_ silmiin* and *pistä_ silmiin* and the range of scores is across all forms (i.e. verb conjugation + noun form). The results are generally poor, except for the idiom *pistää_ silmään/silmiin*. The recall percentages are calculated in table 9 by form and by idiom - all forms match, half of the forms match, at least one of the forms matches (this last measurement most closely matches the one used by Salehi, Cook, and Baldwin in chapter 3.3.2).

Idiom	Pos	Neg	Total	Instances	Score Range
ahmia_silmillään	0	1	1	7	0.3601 - 0.3601
avata_silmänsä	0	8	8	790	0.2106 - 0.3922
iskeä_silmää	1	6	7	368	0.2348 - 0.4200
kääntää_silmät	1	0	1	6	0.4568 - 0.4568
miellyttää_silmää	1	0	1	5	0.4496 - 0.4496
pestä_silmänsä	0	2	2	18	0.2297 - 0.3933
pistää_silmään	6	0	6	130	0.4246 - 0.6666
pistää_silmiin	3	0	3	39	0.5402 - 0.6766
pitää_silmällä	3	19	22	1315	0.1740 - 0.4516
ristiä_silmänsä	0	3	3	76	0.2539 - 0.3202
saada_silmiinsä	1	0	1	5	0.4571 - 0.4571
sattua_silmään	0	2	2	18	0.3855 - 0.3899
ummistaa_silmänsä	0	5	5	393	0.1248 - 0.2810
uskoa_silmiään	0	4	4	89	0.3317 - 0.3956
viehättää_silmää	0	1	1	7	0.4113 - 0.4113

Table 8: Recall values for idioms with 'silmä', score cutoff 0.42

Idiom	Pos	Neg	Total	Recall
By form	16	51	67	23.9 %
All match	5	10	15	33.3 %
At least half match	5	04	15	33.3 %
At least one match	7	8	15	46.7 %

Table 9: Recall percentages for idioms with 'silmä', score cutoff 0.42

Form	Score	Count
matkustajan_silmää	0.5407	5
hävetä_silmänsä	0.5241	7
kirveen_silmään	0.5215	5
immen_silmää	0.5182	5
heitti_silmänsä	0.5174	5
irrottaa_silmiään	0.5162	6
sattui_silmääni	0.5094	5
pisti_silmääni	0.5068	12
luo_silmänsä	0.4991	35
räpäytti_silmäänsä	0.4985	5

Table 10: False positives for idioms with 'silmä', score cutoff 0.42, 10 highest scores shown

When it comes to calculating recall, precision and F1 score (harmonic average), the "By form" counts and percentage are used, because false positives are also counted by form, not per idiom. Table 10 lists some of these false positives, i.e. forms that are classified as idiomatic but are not found in the gold idiom list. The precision based on this list is 16 / 71 = 22.5 %, resulting in an F1 score of 23.2 %. However, many forms could be termed as "false" false positives: in the top 10, two forms have include the possessive suffix *-i*, so they are variants of the idioms *sattua_silmään* and *pistää_silmään*. Additionally, *hävetä_silmänsä* is actually part of the three-word idiom *hävetä silmänsä päästään*:

(11) *hävetä silmänsä päästään*
to be be so ashamed that their eyes fall off
'to be [very] ashamed'

Idiom	Pos	Neg	Total	Instances	Score Range
ahmia_adessive	0	1	1	7	0.3601 - 0.3601
avata_nominative	2	16	18	945	0.2008 - 0.4748
iskeä_partitive	1	6	7	368	0.2348 - 0.4200
kääntää_nominative	2	7	9	242	0.1689 - 0.4636
miellyttää_partitive	1	0	1	5	0.4496 - 0.4496
pestä_nominative	0	4	4	34	0.2297 - 0.3933
pistää_illative	11	0	11	191	0.4246 - 0.6766
pitää_adessive	3	19	22	1315	0.1740 - 0.4516
ristiä_nominative	0	3	3	76	0.2539 - 0.3202
saada_illative	1	0	1	5	0.4571 - 0.4571
sattua_illative	2	2	4	36	0.3855 - 0.5094
seurata_adessive	0	5	5	167	0.2260 - 0.3642
ummistaa_nominative	0	10	10	494	0.1229 - 0.3608
uskoa_partitive	0	7	7	161	0.2888 - 0.3956
viehättää_partitive	0	1	1	7	0.4113 - 0.4113

Table 11: Recall values for idioms with 'silmä', by case, score cutoff 0.42

In tables 11 and 12 analysis is shown based on noun case (see appendix B for some information regarding the grammatical noun cases) and compositionality cutoff of 0.42 for a precision of 32.4 % and F1 score of 26.3 %. The list of false positives is essentially the same as in the base case, except without forms with possessive suffixes. Notable here when compared to table 8 are

the increased counts for some of the forms, plus an additional verb *seurata* whose frequency now meets the threshold. Analysis by case brings a modest improvement - likely due to the fact that forms with a possessive suffix are no longer marked as false positives.

Idiom	Pos	Neg	Total	Recall
By form	23	81	104	22.1 %
All match	3	12	15	20 %
At least half match	4	11	15	26.7 %
At least one match	8	7	15	53.3 %

Table 12: Recall percentages for idioms with 'silmä', by case, score cutoff 0.42

Finally, the F1 scores based on case/form, minimum count and compositionality score cutoff are graphed in figure 4.

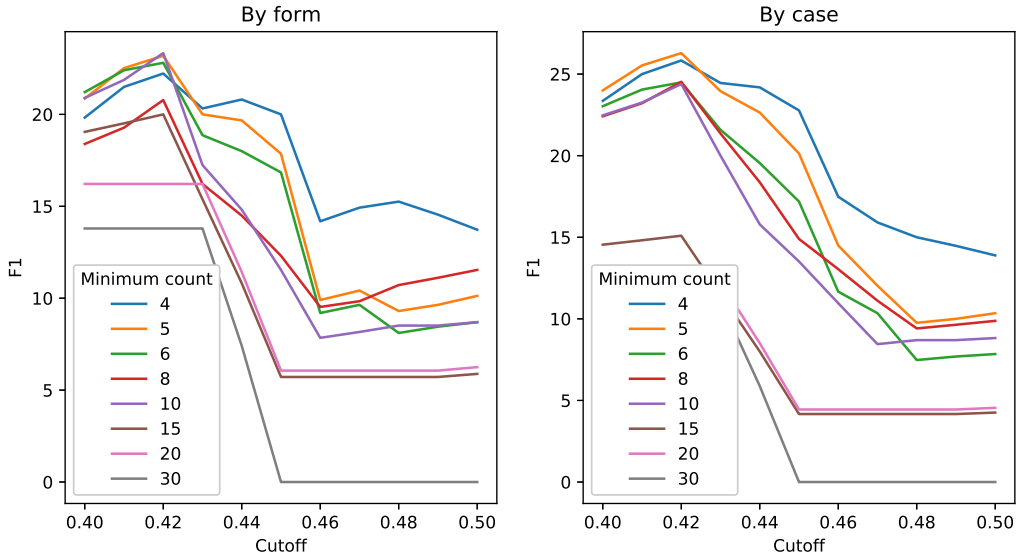


Figure 4: F1 scores by case, frequency and compositionality cutoff

6.4 Evaluation with Lexical Substitution

The true positives and false negatives for the lexical substitution test with minimum count 5 and neighbourhood size 20 are shown in table 13 and table 14.

Idiom	Pos	Neg	Total
ahmia_silmillään	1	0	1
avata_silmänsä	3	5	8
iskeä_silmää	3	4	7
kääntää_silmät	1	0	1
miellyttää_silmää	1	0	1
pestä_silmänsä	0	2	2
pistää_silmään	6	0	6
pistää_silmiin	3	0	3
pitää_silmällä	11	11	22
ristiä_silmänsä	1	2	3
saada_silmiinsä	1	0	1
sattua_silmään	1	1	2
ummistaa_silmänsä	2	3	5
uskoa_silmiään	2	2	4
viehättää_silmää	1	0	1

Table 13: Recall values for idioms with 'silmä', lexical substitution, minimum count 5, neighbourhood size 20

Idiom	Pos	Neg	Total	Recall
By form	37	30	67	55.2 %
All match	7	8	15	46.7 %
At least half match	10	5	15	66.7 %
At least one match	14	1	15	93.3 %

Table 14: Recall percentages for idioms with 'silmä', lexical substitution, minimum count 5, neighbourhood size 20

The list of false positives in table 15 is somewhat different from those listed by the compositionality tests. Precision from 37 true positives and 266 false ones is 12.2 % for an F1 score of 20 %.

Form	Count
silmästä_silmään	315
hieroi_silmiään	127
silmä_silmää	100
sulkien_silmänsä	68
siristi_silmiään	65
hieroi_silmiänsä	47
kuivin_silmin	46
toista_silmäänsä	45
pyyhkii_silmiään	42
räpytteli_silmiään	38
kuivasi_silmänsä	35

Table 15: False positives for idioms with 'silmä', lexical substitution, minimum count 5, neighbourhood size 20, 10 highest counts shown

Table 16 shows a number of examples of the lexical substitutes for the false negatives. For some, a clear substitute is found based on a synonymic verb (*avata* -> *aukaista*, *ummistaa* -> *sulkea*) and some have the noun substituted (*käsi* 'hand', *katse* 'look'). For the idiom *pitää silmällä* a common variant seems to the discontinuous expression *pitää tarkasti silmällä* 'keep a close eye on'. Partial expressions of this kind are marked as false negatives as the methodology cannot account for gappy bigrams.

Form	Lexical substitutes	Score
avasi_silmänsä	aukaisi_silmänsä	0.9045
	nosti_silmänsä	0.8362
iski_silmää	vilkutti_silmää	0.8144
käänsivät_silmänsä	käänsivät_katseensa	0.7737
pitää_silmällä	tarkasti_silmällä	0.7689
pitämään_silmällä	pitämään_tarkasti	0.7468
risti_silmänsä	risti_kätensä	0.7577
seurasi_silmillään	seurasi_katseellaan	0.8065
ummistin_silmäni	suljin_silmäni	0.8539

Table 16: Lexical substitution false negative examples for 'silmä', minimum count 5, neighbourhood size 20

The results based on noun case are shown in table 17. The false positive score is improved at 18.2 % with a slightly better overall performance with F1 score of 27 %.

Idiom	Pos	Neg	Total	Recall
By form	55	49	104	52.9 %
All match	5	10	15	33.3 %
At least half match	9	6	15	60 %
At least one match	15	0	15	100 %

Table 17: Recall percentages for idioms with 'silmä', by case, lexical substitution, minimum count 5, neighbourhood size 20

Finally, the scores based on case/form, minimum count and neighbourhood size is graphed in figure 5. The results with low minimum count cutoff values seem somewhat odd, possibly due to noisiness inherent with dealing with a very small number of token instances.

When using a minimum count cutoff 30 and neighbourhood size 1 (that is, lexical substitution test is only considered successful if the very nearest neighbour has a form that can be considered a lexical substitute), recall is 87.1 % and precision is 39.1 % for an F1 score of 54 %.

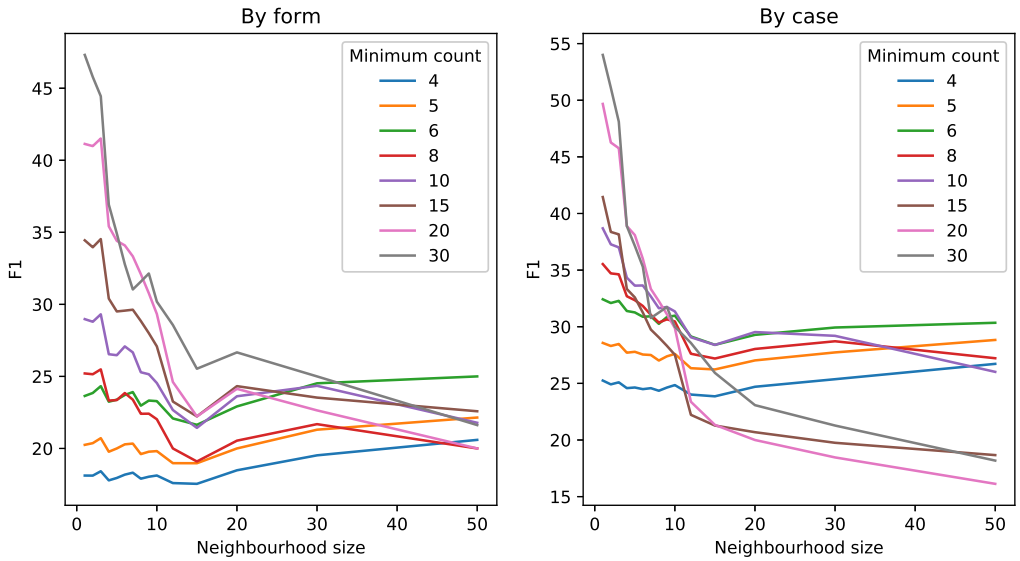


Figure 5: Lexical substitution F1 scores by case, frequency and neighbourhood size

6.5 Subword Embeddings

As seen in tables 18 and 19, the compositionality scores are lower across the board for fastText. This is to be expected, as with subword embeddings forms with common substrings are closer to each other (i.e. *pistää_silmään* is much more similar to the combination of vectors *pistää* and *silmään* when compared to base word2vec). There also seems to be an inverse correlation with the score and the bigram length.

Most compositional			Least compositional		
Form	Score	Count	Form	Score	Count
kastehelmet_kimaltelivat	0.0232	11	koko_ajan	0.4566	3740
riisui_päällystakkinsa	0.0292	13	ajan_tapaa	0.4241	11
nuku_nurmilintu	0.0299	12	näin_koko	0.4205	80
kulmakarvat_varjostivat	0.0312	10	toista_kertaa	0.4141	527
kulmakarvat_vetäytyivät	0.0318	13	hädän_tullen	0.4087	39

Table 18: Most and least compositional bigrams for fastText, minimum count 10

Most compositional			Least compositional		
Form	Score	Count	Form	Score	Count
silmät_muljahtelivat	0.040	3	silmä_kantoi	0.383	222
silmät_rävähtivät	0.041	13	silmä_kantaa	0.348	109
silmillään_tuijottaen	0.043	5	silmä_kanna	0.347	4
silmin_tuijottamaan	0.043	8	etsi_silmä	0.339	3
silmänsä_rävähtivät	0.044	4	juuri_silmän	0.339	3

Table 19: Most and least compositional bigrams for V+N / N+V, N=*silmä* 'eye', fastText version

The recall values are much better for fastText when compared the base case, as seen in tables 20 and 21 for minimum token frequency of 30 and compositionality score threshold of 0.14 (again, as the best performance is achieved with this cutoff value). Precision for this is 46.5 % from 20 true and 23 false positives, resulting in an F1 score of 54.1 %.

Idiom	Pos	Neg	Total	Instances	Score Range
avata_nominative	3	2	5	807	0.1159 - 0.1773
iskeä_partitive	3	1	4	342	0.1368 - 0.1901
kääntää_nominative	2	0	2	156	0.1410 - 0.1421
pistää_illative	2	0	2	92	0.1786 - 0.1855
pitää_adessive	7	3	10	1159	0.1185 - 0.2054
ristiä_nominative	1	0	1	47	0.1764 - 0.1764
seurata_adessive	0	2	2	133	0.0969 - 0.0982
ummistaa_nominative	0	3	3	393	0.0861 - 0.1033
uskoa_partitive	2	0	2	110	0.1413 - 0.1767

Table 20: Recall values for idioms with 'silmä', by case, score cutoff 0.14, minimum count 30, using fastText

Idiom	Pos	Neg	Total	Recall
By form	20	11	31	64.5 %
All match	4	5	9	44.4 %
At least half match	7	2	9	77.8 %
At least one match	7	2	9	77.8 %

Table 21: Recall percentages for idioms with 'silmä', by case, score cutoff 0.14, minimum count 30, using fastText

The effect of subwords embeddings is quite dramatic, as shown in figure 6. Even with low minimum counts, the results are better than in the base case and - unlike in figure 4 on page 46 - the performance improves when frequency cutoff is increased.

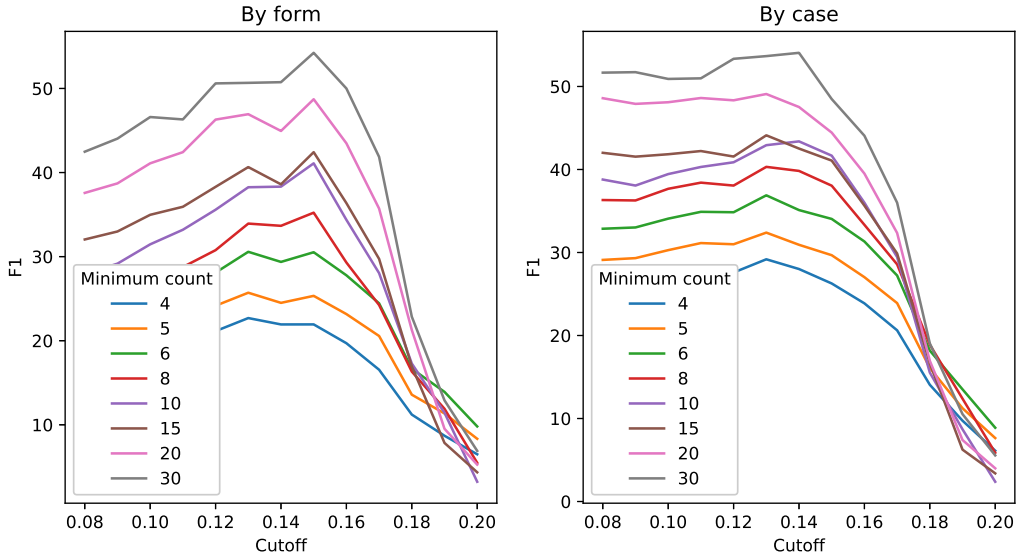


Figure 6: fastText compositionality F1 scores by case, frequency and compositionality cutoff

As with the compositionality test, the lexical substitution test shows improved performance when compared to the base case. The model is also more well-behaved regarding the frequency cutoff. The results are shown in figure 7. The recall percentages for neighbourhood size 1 and minimum count 30 are shown in table 22. The precision with these parameters is 41 %, yielding an F1 score of 45.7 %.

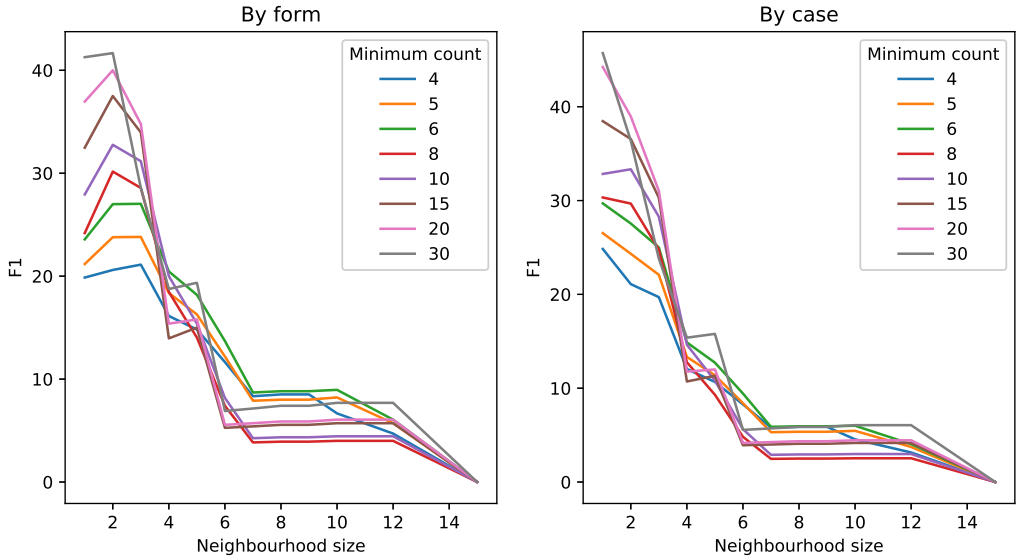


Figure 7: fastText lexical substitution F1 scores by case, frequency and neighbourhood size

Idiom	Pos	Neg	Total	Recall
By form	16	15	31	51.6 %
All match	2	7	9	22.2 %
At least half match	6	3	9	66.7 %
At least one match	6	3	9	66.7 %

Table 22: Lexical substitution recall percentages for idioms with 'silmä', by case, neighbourhood size 1, minimum count 30, using fastText

6.6 Classification

The verbs to be classified with statistics regarding their labels are shown in table 23. Table 24 shows the results from the classification runs for various parameters and options. The window size refers to the size of neighbourhood when calculating the context embedding, not the one used for training the model (which always uses a window size of 5). The Test column shows the best accuracy for the various epochs tried and the training accuracy is for that run. In all cases, using exponential decay for context calculation improves the performance.

Idiom	Non-literal	Literal
ahmia_adessive	17	0
avata_nominative	56	738
iskeä_partitive	343	1
kääntää_nominative	20	268
miellyttää_partitive	10	0
pestä_nominative	1	61
pistää_illative	172	6
pitää_adessive	1252	0
ristiä_nominative	63	2
saada_illative	4	3
sattua_illative	48	2
seurata_adessive	190	0
ummistaa_nominative	97	326
uskoa_partitive	102	0
viehättää_partitive	15	0
Total	2400	1407

Table 23: Classification labels for verbs

Model	Context	Window size	Train	Test
word2vec	Default	5	88.82	82.97
word2vec	Default	10	89.80	82.17
word2vec	Exponential	5	91.14	84.24
word2vec	Exponential	10	94.02	84.08
fastText	Default	5	89.05	83.15
fastText	Default	10	88.90	80.04
fastText	Exponential	5	88.93	83.24
fastText	Exponential	10	88.16	83.70

Table 24: Classification results for various options

As a side note, while labelling the data set, an idiom was found that had three distinct senses, as shown in example 12. This echoes the sentiment from chapter 2.5: it is possible for MWEs to have more than two senses.

(12) *ummistaa silmänsä*
to close one's eyes

'to choose to not see something [bad]'

'to die'

6.7 Exploration: Odds and Ends

6.7.1 The Curious Case of Noun Case and Minimal Pairs

There were few idiomatic minimal pairs in the gold data and none that had both forms represented in the Gutenberg test data. However, one pair was found when labelling the data for classification:

(13) *iskeä* *silmää*
to punch the eye
'wink'

(14) *iskeä* *silmänsä*
to punch [one's] eye
'to become interested in something'

In table 25, the in-group similarity average between nominative forms is 0.54 and for partitive forms is 0.46, while the out-group similarity average between the forms in different cases is 0.38. This means that nominative forms are closer to other nominative forms than partitive forms, and the same holds for partitive forms. This is in line with the expectation that noun case is important.

Case		
	nominative	partitive
nominative (silmänsä)	**0.54**	0.38
partitive (silmää)		**0.46**

Table 25: Iskeä silmää vs iskeä silmänsä by noun case

The results for groupings based on verb mood and tense are shown tables 26 and 27, which do not show a strong in-group/out-group effect.

Mood		
	indicative	infinitive
indicative	**0.40**	0.43
infinitive		**0.52**

Table 26: Iskeä silmää vs iskeä silmänsä by verb mood

Tense			
	past_imperfective	present_simple	undefined
past_imperfective	**0.43**	0.38	0.45
present_simple		**0.44**	0.41
undefined			**0.52**

Table 27: Iskeä silmää vs iskeä silmänsä by verb tense

Tables 28 and 29[40] show the scores by case for the idioms *pitää silmällä* and *pistää silmään* with similar results to table 25. Only the adessive and illative cases, respectively, are idiomatic, the rest are literal.

Case			
	adessive	nominative	partitive
adessive	**0.47**	0.35	0.35
nominative		**0.53**	0.52
partitive			**0.57**

Table 28: Pitää silmällä by noun case

Case			
	illative	nominative	partitive
illative	**0.47**	0.38	0.42
nominative		**0.55**	0.34
partitive			**0.43**

Table 29: Pistää silmään by noun case

[40] The relatively high similarity score between illative and partitive cases may be due to the fact that the form *silmääni* can be analysed as belonging to either case, in which case they are counted in both groups.

6.7.2 Idiomatic Synonymy

In table 30 the most common substitutes for the idiom *tehdä mieli* from a neighbourhood of 20 are shown. The substitutes are classified as: closest unigram, lexical substitution (one component has a different lemma), both have same lemmas, lexical substitution with inverted word order, same lemmas in different order and any bigram not included in the other categories.

	teki_mieli (count 1020)	*tekisi_mieli* (369)	*tekee_mieli* (292)
Unigram	halutti (0.7342) täytyi (0.7105)	sopisi (0.7281) haluttaisi (0.6993) pitäisi (0.6801)	pitänee (0.7078) täytyy (0.6947)
Lex sub			
Same lemmas	teki_mielensä (0.8459) tekisi mieli (0.8380) tekee mieli (0.8015)	tekee_mieli (0.8919) tekisi_mieleni (0.8514) teki_mieli (0.8380)	tekisi_mieli (0.8919) tekisi_mieleni (0.8191) teki_mieli (0.8015)
Inverse lexsub	mieli_hiukan (0.7160)	mieli_hiukan (0.6895)	
Inverse same	mieli_teki (0.7674) mielensä_teki (0.7354)	mieleni_tekisi (0.7034)	mieli_tekee (0.7823) mieleni_tekisi (0.7221) mieleni_tekee (0.7087)
Any bigram	tuli_halu (0.7831) häntä_halutti (0.7753) silloin_täytyi (0.7502)	täytyy_väkisinkin (0.7008) oikeastaan_pitäisi (0.6927) tuli_halu (0.6837)	minäkin_tahdon (0.7252) juuri_täytyy (0.7222) silloin_täytyy (0.7221)

Table 30: Tehdä mieli substitutions, neighbourhood size 20

None of the forms have direct lexical substitutes, in which case these forms would be classified as idiomatic. All neighbours relate to either desire or the need / obligation to do something.

As could be expected, *mieli tehdä* variants are close to the meaning of *tehdä mieli*, but it's unclear whether this would be a case of true idiomatic synonymy or whether these should be considered as variants of the same idiomatic construction.

6.8 Clustering

The results for the clustering the idiom *ummistaa silmänsä* are shown in table 31. Cutoff value of 0.073 is used here, as it provided the best performance for this idiom, although it is not optimal against the whole idiom list.

The results here are mixed. The clustering method does actually do a fair job separating the instances based on the compositionality scores, but the instance counts for the literal and non-literal clusters (based on the compositionality scores) differ quite a lot from the classification gold labels in table 23 on page 54: here there are 210 and 159 "non-literal" and "literal" instances, respectively,[41] while in the gold data the numbers are 97 and 326. Other than checking the counts, no comparison was made between the clustered instances and the gold list.

Whether the good result for compositionality score clustering is purely accidental or whether it reflects real potential of the method cannot be determined based on this one example alone.

Idiom	Pos	Neg	Total	Instances	Score Range
ummistaa_nominative:1	9	0	9	210	0.0733 - 0.1524
ummistaa_nominative:2	1	2	3	159	0.0634 - 0.0772

Table 31: Recall values for 'ummistaa silmänsä' clusters, by case, score cutoff 0.073, minimum count 5, using fastText

6.9 Analysis and Discussion

To summarise the idiomaticity tests, table 32 lists the recall, precision and F1 scores for various compositionality score cutoff / neighbourhood size values.

As the results in the previous chapters show, the compositionality and lexical substitution tests have some value, but the precision and recall values leave something to be desired, so neither is likely to be useful in isolation. Analysing the idioms by case yields a small improvement. Recall is a little bit lower - probably because additional forms that are identified as being part of the idiom are less frequent and thus more noisy, but this is more than made up fewer false positives.

The compositionality scores are much lower for fastText, but the general

[41] As some forms fail to meet the frequency cutoff threshold after the clustering, the total count is lower.

Algorithm	Test	Form	Cutoff	Neighb. size	Min	Recall	Precision	F1
word2vec	compositionality	form	0.42		5	23.9	22.5	23.2
word2vec	compositionality	case	0.42		5	22.1	32.4	26.3
word2vec	substitution	form		20	5	55.2	12.2	20
word2vec	substitution	case		20	5	52.9	18.2	27
word2vec	substitution	case		1	30	87.1	39.1	54
fastText	compositionality	case	0.14		30	64.5	46.5	54.1
fastText	substitution	case		1	30	51.6	41	45.7

Table 32: Recall, precision and F1 scores for 'silmä'

performance is better, and the model is much more well-behaved when the minimum token frequency is increased. The superior performance is even more notable when considering the bias regarding bigram length.

Part of the poor performance in general may be due to the fact that (as noted in chapter 2.6) idiom-prone verbs are generally semantically neutral and highly polysemous. For example, the idiom *pitää silmällä* is always idiomatic as no plausible literal interpretation exists, yet the compositionality scores are low.

For the lexical substitution test, the best results are achieved when only the nearest neighbour is taken into account. The original algorithm works a little bit better than fastText for this test. Lexical substitution also seems to have some use for finding idiomatic synonyms.

Regarding the importance of grammatical case in chapter 6.7, the results are consistent (or at least, not contradictory) with the findings of Nenonen from chapter 2.6. However, these should not be considered as any kind of robust statistical evidence.

The classification results based on the small labelled set fail to go above 90 %. This may be due to many factors, not least of which is the low amount of data.

Based on the small clustering test, clustering seems to be have some use for separating literal and non-literal interpretations from each other, at least when measured by the compositionality score. The specific example, however is cherry-picked and relies on a priori knowledge regarding the idiom - that is, that literal and non-literal interpretations exist in relatively balanced proportions. It would make little sense to cluster an idiom that has no literal intrepretation. The approach therefore is not applicable as a general solution for distinguishing between literal and non-literal interpretations, or at least not without major improvements.

7 Conclusions

As is evident from the previous chapter, the results from the various experiments are mixed. In retrospect, MWE classification and disambiguation is such a hard topic that it would have been unrealistic to expect great results, at least with such simple methods.

It is clear that compositionality and lexical substitution tests are not silver bullets. The question here is: how much of the lack of performance is due to deficiencies in the methodologies, and how much due to inherent semantic properties? This remains unclear. As noted in chapter 2.3, the only common criterion for an idiom seem to be conventionality and many idioms are actually relatively transparent and/or compositional. It also doesn't help that current methodology conflates multiple senses for both the idiom itself and its components.

Is compositionality a useful feature for quantifying idiomaticity? In a way, testing compositionality is a "classical" approach with a long history in the literature. The fact that F1 scores for fastText edge above 50 % (see chapter 6.5) suggest that there the method does still have some legs. It is also possible that the method does capture some degree of compositionality.

Some of the lackluster performance may be due to preprocessing (see chapter 5.1). The identification of verbal idioms has been done with a very simple method which does not account for discontinuous expressions and also misidentified various other expressions as idiomatic. Adding proper syntactic parsing (either based on constituency or dependency parsing) and handling discontinuous MWEs could improve the results.

There is also a certain mismatch between the data set and the gold idiom list. While there is no exact dating for the Gutenberg data, the newest works are generally from the 19th or early 20th century, while the gold idiom list is from late 20th century. This means that there are idioms in the data set that do not exist in the gold data, which will certainly increase the false positive rate. Many idioms in the gold list do not occur in the data as they are too new. The older data also contains archaic and colloquial word forms (see 4.1); accounting for these might increase the accuracy of the methods.Despite various improvements that could be done with data sets and preprocessing, the major impediment to performance remains handling multiple senses. Clustering only the idiom seems insufficient, thus one would need to account for the polysemy of the components (as was done 3.2.2), potentially leading to an explosion of forms - the number of which, per idiom, is already quite high.

Handling these with simple embeddings could therefore become cumbersome.

Embeddings seem to work fairly well for determining whether grammatical case is more important for indexing idiomatic meaning than other grammatical properties, although, as already noted, the results based on a limited test do not rise to the level of statistical validity. The methods also seem to have some use finding cases of idiomatic synonymy (or paraphrases).

To circle back to the first question: how good are embeddings in general for analysis Finnish idioms? Based on the various experiments one can say that the results are mixed. Subword embeddings work better for some tasks, such as quantifying idiomaticity, whereas the base model is better for tasks based on comparing similarity to other tokens like lexical substitution test and idiomatic synonymy.

Despite various improvements that could be done with data sets and preprocessing, the major impediment to performance (at least for compositionality) remains handling multiple senses. Clustering only the idiom seems insufficient, thus one would need to account for the polysemy of the components (as was done 3.2.2), potentially leading to an explosion of forms - the number of which, per idiom, is already quite high. Handling these with simple embeddings could therefore become cumbersome.

8 Future Considerations

There are many things that could have been done better in this thesis. First of all, the corpus could have been chosen better. A modern corpus, such as Suomi24 (Aller Media ltd., 2014) would be more consistent with the gold idiom set. It would also mean that one would not need to care about archaic forms, although it would probably still add colloquial word forms that might not be recognised by the current methodology. It would also be possible to use a variant of omorfi that can handle these archaic forms.[42] Word embeddings themselves could also be used to normalise the archaic forms as was done in (Hämäläinen and Hengchen, 2019). The normalisation approach would have the benefit of a reduced vocabulary leading to higher quality embeddings.

The false positives identified by both tests contained some expressions that are actually valid verbal idioms. These could manually curated and added to the gold list, thus improving the scores a little bit.

The analysis would also be more accurate if syntax and morphology were taken properly into account. This would mean parsing the data with modern methodology such as Turku NLP parser[43], Turku neural parser[44], FinBERT[45] or FinnPos[46]. Dependency parsing would be preferable since it could naturally take discontinuous expressions and variations of word order into account. This would also allow for easier analysis of longer expressions, not just bigrams, and also enable taking advantage of various features obtained through the parsing, such as part of speech tags. Quantifying the prevalence of variations from the "canonical" form could also make it possible to study correlations between productivity and transparency/figuration, as hypothesised in (Sheinfux et al., 2019).

Regarding embeddings, it would be interesting to see if the bias of subword embeddings with expression length has to do with the actual algorithm. In any case, normalising based on bigram length might yield improved performance. Using morfessor (which is cognitively motivated) could also yield some improvement over fastText. Using the whole sentence as a context - as in the SentVec algorithm (Salton, Ross, and Kelleher, 2016) - could also improve the performance for this task. Combining the compositionality and lexical substitution tests to a unified measure might also be useful. The tests

[42] https://github.com/jiemakel/omorfi
[43] https://turkunlp.org/finnish_nlp.html
[44] https://github.com/TurkuNLP/Turku-neural-parser-pipeline
[45] https://turkunlp.org/FinBERT/
[46] https://github.com/mpsilfve/FinnPos

could also be used as part of a larger idiomaticity testing mechanism.

The main avenue for improvement, however, would be handling multiple senses. This proved to be cumbersome in the current approach, where only simple embeddings have been used. The state of the art has moved on to deep learning models such as ELMo (Peters, Neumann, et al., 2018) and BERT (Devlin et al., 2019) that can handle the different contexts and multiple senses in a more streamlined fashion. This would, of course, be at the cost of additional unavoidable computational complexity.

The diachronic aspect, that is, investigating semantic change of idioms was dropped from scope of the thesis (as outlined in chapter 1.2). One of reasons was the expectation that the kinds of idioms that are studied here do not change meaning (easily or at all), as explained in chapter 2.8.1. On the other hand, it seems difficult to take this statement at face value, that is, that idioms never change. As far as I can tell, this kind of thing has not be not been studied before, which suggests one possible avenue of investigation. Another possibility is that - assuming that idioms don't change - they could be use as "anchors" (fixed points) when comparing corpora across in different time periods.

Even if it might not be possible to study the meaning in change, what can be studied, though, is the relative proportion of literal vs figurative interpretations. For example, this thesis was finalised during the 2020 corona virus outbreak, during which we were repeatedly and emphatically told to wash our hands (*pestä kädet/kätensä*). Thus, the expectation is that the during this time the literal interpretation of the idiom would be much more prevalent when compared to the figurative one.

Aspects of productivity could also be studied, as in the Mumin troll study referred to in chapter 2.2. Would the variation of an idiomatic construction increase as time goes by?

The study of the importance of grammatical case from chapter 2.6 (versus other grammatical properties, such as verb mood and tense) ended up being rather anecdotal. A more through investigation would be required to quantitatively affirm Nenonen's results. There would also be many things regarding idiomatic synonymy that could be studied through embeddings.

References

Aller Media ltd. (2014). *The Suomi 24 Corpus (2016H2)*. URL: http://urn.fi/urn:nbn:fi:lb-2017021506.

Anastasiou, Dimitra et al., eds. (2009). *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*. Association for Computational Linguistics. URL: https://dl.acm.org/citation.cfm?id=1698239.

Bartunov, Sergey et al. (2016). "Breaking Sticks and Ambiguities with Adaptive Skip-gram". In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Ed. by Arthur Gretton and Christian C. Robert. Vol. 51, pp. 130–138. URL: http://proceedings.mlr.press/v51/bartunov16.html.

Bender, Emily M. (2019). "English isn't generic for language, despite what NLP papers might lead you to believe". In: *Symposium on Data Science & Statistics*. URL: https://faculty.washington.edu/ebender/papers/Bender-SDSS-2019.pdf.

Biber, Douglas and Susan Conrad (2009). *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press. DOI: 10.1017/CBO9780511814358.

Birke, Julia and Anoop Sarkar (2006). "A Clustering Approach for the Nearly Unsupervised Recognition of Nonliteral Language". In: *Proceedings of EACL-06*. School of Computing Science, Simon Fraser University. URL: https://www.aclweb.org/anthology/E06-1042.

Bojanowski, Piotr et al. (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics*. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/Q17-1010/.

Bowern, Claire L. (2019). "Semantic Change and Semantic Stability: Variation is Key". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pp. 48–55. DOI: 10.18653/v1/W19-4706. URL: https://www.aclweb.org/anthology/W19-4706.

Butt, Miriam Jessica (2010). *The Light Verb Jungle: Still Hacking Away*. URL: http://ling.uni-konstanz.de/pages/home/butt/main/papers/cp-volume.pdf.

Caselles-Dupré, Hugo, Florian Lesaint, and Jimena Royo-Letelier (2018). *Word2vec applied to Recommendation: Hyperparameters Matter*. URL: https://doi.org/10.1145/3240323.3240377.

Cilibrasi, Rudi and Paul M. B. Vitányi (2007). "The Google Similarity Distance". In: *IEEE Transactions on Knowledge and Data Engineering* abs/cs/0412098, pp. 370–383. DOI: 10.1109/TKDE.2007.48. URL: https://ieeexplore.ieee.org/abstract/document/4072748.

Colson, Jean-Pierre (2017). "The IdiomSearch Experiment: Extracting Phraseology from a Probabilistic Network of Constructions". In: *Computational and Corpus-Based Phraseology, Second International Conference, Europhras 2017 London, UK, November 13–14, 2017, Proceedings*. Springer, Cham, pp. 16–28. URL: https://doi.org/10.1007/978-3-319-69805-2_22.

Conklin, Kathy and Norbert Schmitt (2012). "The processing of formulaic language". In: *Annual Review of Applied Linguistics* 32, pp. 45–61. URL: https://doi.org/10.1017/S0267190512000074.

Cook, Paul, Afsaneh Fazly, and Suzanne Stevenson (2007). "Pulling their Weight: Exploiting Syntactic Forms for the Automatic Identification of Idiomatic Expressions in Context". In: *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*. Association for Computational Linguistics, pp. 41–48. URL: http://www.cs.toronto.edu/~pcook/CFS2007.pdf.

Creutz, Mathias and Krista Lagus (2005). *Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0*. Tech. rep. Helsinki University of Technology.

Daudaravičius, Vidas and Ruta Murcinkevičiene (2004). "Gravity counts for the boundaries of collocations". In: *International Journal of Corpus Linguistics* 9.2, pp. 321–348. DOI: 10.1075/ijcl.9.2.08dau. URL: https://doi.org/10.1075/ijcl.9.2.08dau.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Dhar, Prajit, Janis Pagel, and Lonneke van der Plas (2019). "Measuring the Compositionality of Noun-Noun Compounds over Time". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pp. 234–239. DOI: 10.18653/v1/W19-4729. URL: https://www.aclweb.org/anthology/W19-4729.

Döbrössy, Bálint et al. (2019). "Investigating sub-word embedding strategies for the morphologically rich and free phrase-order Hungarian". In: *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. Association for Computational Linguistic, pp. 187–193. URL: https://www.aclweb.org/anthology/W19-4321.

Dubossarsky, Haim, Eitan Grossman, and Daphna Weinshall (2017). "Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1136–1145. DOI: 10.18653/v1/D17-1118. URL: https://aclweb.org/anthology/papers/D/D17/D17-1118/.

Dyvik, Helge (2004). "Translations as Semantic Mirrors: From Parallel Corpus to Wordnet". In: *Language and Computers* 49, pp. 311–326. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.149.2412.

Fazly, Afsaneh, Paul Cook, and Suzanne Stevenson (2009). "Unsupervised Type and Token Identification of Idiomatic Expressions". In: *Computational Linguistics* 35, pp. 61–103. DOI: 10.1162/coli.08-010-R1-07-048. URL: https://www.mitpressjournals.org/doi/abs/10.1162/coli.08-010-R1-07-048.

Firth, John Rupert (1957). ""A Synopsis of Linguistic Theory 1930-1955" in Studies in Linguistic Analysis". In: *The Philological Society*.

Garcia, Marcos (2018). "Comparing bilingual word embeddings to translation dictionaries for extracting multilingual collocation equivalents". In: *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*.

Geeraert, Kristina, R. Harald Baayen, and John Newman (2018). ""Spilling the bag" on idiomatic variation". In: *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*. URL: https://research.monash.edu/en/publications/spilling-the-bag-on-idiomatic-variation.

Goldberg, Yoav (2017). *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers.

Gries, Stefan (2009). *Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora*. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.158.71.

— (2010). "Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora". In: *Proceedings of Corpus Linguistics 2009*.

Hakala, Tero et al. (2018). "Information properties of morphologically complex words modulate brain activity during word reading". In: *Human Brain Mapping* 39.6, pp. 2583–2595. URL: https://doi.org/10.1002/hbm.24025.

Hall, Timothy (2009). "The Fossilization-Formula Interface". In: *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics* 9.2. URL: https://journals.cdrs.columbia.edu/wp-content/uploads/sites/12/2015/06/3.6-Hall-2009.pdf.

Hämäläinen, Mika and Simon Hengchen (2019). "From the Paft to the Fiiture: a Fully Automatic NMT and Word Embeddings Method for OCR Post-Correction". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 431–436. DOI: 10.26615/978-954-452-056-4_051. URL: https://www.aclweb.org/anthology/R19-1051/.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky (2016). *Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change*. URL: https://aclweb.org/anthology/D16-1229.

Hoang, Hung Huu, Su Nam Kim, and Min-Yen Kan (2009). "A Re-examination of Lexical Association Measures". In: *2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation, Applications: Proceedings of the Workshop*, pp. 23–30. URL: https://www.aclweb.org/anthology/W/W09/W09-29.

Huang, Eric H. et al. (2012). "Improving Word Representations via Global Context and Multiple Word Prototypes". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. ACL '12. Jeju Island, Korea: Association for Computational Linguistics, pp. 873–882. URL: http://dl.acm.org/citation.cfm?id=2390524.2390645.

Hurwitz, Daniel (2012). "Morphological and Lexical Decomposition as a Basis for Identifying Multiword Expressions". MA thesis. Technion - Israel Institute of Technology.

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli (2016). "Embeddings for Word Sense Disambiguation: An Evaluation Study". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 897–907. URL: https://www.aclweb.org/anthology/P16-1085.

Jurafsky, Daniel and James H. Martin (2018). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition: 3rd Edition draft*. URL: https://web.stanford.edu/~jurafsky/slp3/.

Katz, Graham and Eugenie Giesbrecht (2006). "Automatic Identification of Non-Compositional Multi-Word Expressions using Latent Semantic Analysis". In: *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. Association for Computational Linguistics, pp. 12–19. URL: https://aclweb.org/anthology/W06-1203.

Kettunen, Kimmo, Tuula Pääkkönen, and Mika Koistinen (2016). "Kansalliskirjaston digitoitu historiallinen lehtiaineisto 1771–1910: sanatason laatu, kokoelmien käyttö ja laadun parantaminen." In: *Informaatiotutkimus* 35.3. URL: https://journal.fi/inf/article/view/59433.

Keysar, Boaz and Bridget Bly (1995). "Intuitions of the Transparency of Idioms: Can One Keep a Secret by Spilling the Beans". In: *Journal of Memory and Language* 34.1, pp. 89–109. URL: https://doi.org/10.1006/jmla.1995.1005.

Kortelainen, Kristiina (2012). "Ei ole kaikki muumit laaksossa. Tutkimus suomen kielen idiomikonstruktion produktiivisuudesta". MA thesis. University of Turku. URL: https://www.utupub.fi/handle/10024/73976.

Koskenniemi, Kimmo, Georg Rehm, and Hans Uszkoreit, eds. (2012). *The Finnish Language in the Digital Age*. Springer. URL: http://dx.doi.org.portal.lib.fit.edu/10.1007/978-3-642-27248-6.

Le, Quoc and Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31st International Conference on Machine Learning*, vol. 32. URL: https://ai.google/research/pubs/pub44930.

Leminen, Alina, Sini Jakonen, et al. (2016). "Neural mechanisms underlying word- and phrase-level morphological parsing". In: *Journal of Neurolinguistics* 38, pp. 26–41. URL: https://doi.org/10.1016/j.jneuroling.2015.10.003.

Leminen, Alina, Eva Smolka, et al. (2018). "Morphological processing in the brain: The good (inflection), the bad (derivation) and the ugly (compounding)". In: *Cortex*. URL: https://doi.org/10.1016/j.cortex.2018.08.016.

Lichte, Timm et al. (2019). "Lexical encoding formats for multi-word expressions: The challenge of 'irregular' regularities". In: *Representation and parsing of multiword expressions: Current trends*. Ed. by Yannick Parmentier and Jakub Waszczuk. Language Science Press, pp. 1–33. URL: https://langsci-press.org/catalog/book/202.

Louwerse, Max et al. (2004). "Variation in Language and Cohesion across Written and Spoken Registers". In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 26. URL: https://escholarship.org/uc/item/7d8631cr.

Markantonatou, Stella et al., eds. (2018). *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*. Language Science Press. URL: http://langsci-press/catalog/book/204.

Melamud, Oren, Jacob Goldberger, and Ido Dagan (2016). "context2vec: Learning Generic Context Embedding with Bidirectional LSTM". In: pp. 51–61. DOI: 10.18653/v1/K16-1006. URL: https://aclweb.org/anthology/K/K16/K16-1006.

Mikolov, Tomas, Kai Chen, et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *Proceedings of the International Conference on Learning Representations (ICLR 2013)*. URL: https://arxiv.org/abs/1301.3781.

Mikolov, Tomas, Ilya Sutskever, et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pp. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.

Milburn, Trudy (2015). "Speech Community". In: *The International Encyclopedia of Language and Social Interaction*. The Wiley Blackwell-ICA

Encyclopedias of Communication. Wiley. DOI: 10.1002/9781118611463.wbielsi040. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118611463.

Mitkov, Ruslav, ed. (2017). *Computational and Corpus-Based Phraseology, Second International Conference, Europhras 2017 London, UK, November 13–14, 2017, Proceedings*. Springer. URL: https://doi.org/10.1007/978-3-319-69805-2_22.

Moirón, Begoña Villada and Jörg Tiedemann (2006). "Identifying idiomatic expressions using automatic word-alignment". In: *Proceedings of the EACL 2006 Workshop on Multi-word expressions in a multilingual context*, pp. 33–40. URL: https://aclweb.org/anthology/W/W06/W06-2405.

Moreau, Erwan et al. (2018). "Semantic reranking of CRF label sequences for verbal multiword expression identification". In: *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*. Language Science Press, pp. 177–207. URL: http://langsci-press.org/catalog/book/204.

Neelakantan, Arvind et al. (2015). "Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 1059–1069. DOI: 10.3115/v1/D14-1113. URL: https://www.aclweb.org/anthology/D14-1113/.

Nenonen, Marja (2002a). "Helppo nakki?" In: *Virittäjä* 3. URL: https://journal.fi/virittaja/article/download/40199/9626.

— (2002b). "Idiomit ja leksikko Lausekeidiomien syntaktisia, semanttisia ja morfologisia piirteitä suomen kielessä". PhD thesis. Joensuun Yliopisto.

— (2007). "Prototypical Idioms: Evidence from Finnish". In: *SKY Journal of Linguistics* 20, pp. 309–330. URL: http://www.linguistics.fi/julkaisut/SKY2007/NENONEN.pdf.

Niemi, Jussi et al. (2013). "Idiomatic proclivity and literality of meaning in body-part nouns: Corpus studies of English, German, Swedish, Russian and Finnish". In: *Folia Linguistica* 47. DOI: 10.1515/flin.2013.009. URL: https://www.researchgate.net/publication/273930399_Idiomatic_proclivity_and_literality_of_meaning_in_body-part_nouns_Corpus_studies_of_English_German_Swedish_Russian_and_Finnish.

Nunberg, Geoffrey, Ivan A. Sag, and Thomas Wasow (1994). "Idioms". In: *Language* 70.3, pp. 491–538. URL: http://lingo.stanford.edu/sag/papers/idioms.pdf.

Parizi, Ali Hakimi and Paul Cook (2018). "Do Character-Level Neural Network Language Models Capture Knowledge of Multiword Expression Compositionality?" In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, pp. 185–192. URL: https://aclweb.org/anthology/W18-4920.

Parmentier, Yannick and Jakub Waszczuk, eds. (2019). *Representation and parsing of multiword expressions: Current trends*. Language Science Press. URL: https://langsci-press.org/catalog/book/202.

Peng, Jing, Anna Feldman, and Ekaterina Vylomova (2014). *Classifying Idiomatic and Literal Expressions Using Topic Models and Intensity of Emotions*. URL: https://www.aclweb.org/anthology/D14-1216.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

Peters, Matthew E., Waleed Ammar, et al. (2017). "Semi-supervised sequence tagging with bidirectional language models". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. URL: https://doi.org/10.18653/v1/P17-1161.

Peters, Matthew E., Mark Neumann, et al. (2018). "Deep contextualized word representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2227–2237. URL: https://www.aclweb.org/anthology/N18-1202/.

Petrova, Oksana (2011). "Of Pearls and Pigs: A Conceptual-Semantic Tiernet Approach to Formal Representation of Structure and Variation of Phraseological Units". PhD thesis. Åbo Akademi. URL: https://www.doria.fi/handle/10024/69852.

Reisinger, Joseph and Raymond J. Mooney (2010). "Multi-Prototype Vector-Space Models of Word Meaning." In: *Proceedings of the 2010 Annual Conference of the NAACL*.

Sag, Ivan A. et al. (2002). "Multiword Expressions: A Pain in the Neck for NLP". In: *CICLing '02 Proceedings of the Third International Conference*

on Computational Linguistics and Intelligent Text Processing, pp. 1–15. URL: https://dl.acm.org/citation.cfm?id=724004.

Säily, Tanja (2014). "Sociolinguistic variation in English derivational productivity : Studies and methods in diachronic corpus linguistics". PhD thesis. University of Helsinki. URL: https://helda.helsinki.fi/handle/10138/136128.

Salehi, Bahar, Paul Cook, and Timothy Baldwin (2015). "A Word Embedding Approach to Predicting the Compositionality of Multiword Expressions". In: *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pp. 977–983. URL: https://www.aclweb.org/anthology/N15-1099.

— (2018). "Exploiting multilingual lexical resources to predict MWE compositionality". In: *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*.

Salle, Alexandre and Aline Villavicencio (2018). "Incorporating Subword Information into Matrix Factorization Word Embedding". In: *Proceedings of the Second Workshop on Subword/Character LEvel Models*. Association for Computational Linguistics, pp. 66–71. URL: https://www.aclweb.org/anthology/W18-1209.

Salton, Giancarlo D., Robert J. Ross, and John D. Kelleher (2016). "Idiom Token Classification using Sentential Distributed Semantics". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL '16. Berlin, Germany: Association for Computational Linguistics, pp. 194–204. URL: https://aclweb.org/anthology/P16-1019.

Savary, Agata, Marie Candito, et al. (2018). "PARSEME multilingual corpus of verbal multiword expressions". In: *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pp. 87–147. URL: https://repository.ubn.ru.nl/handle/2066/200301.

Savary, Agata, Carlos Ramisch, et al. (2017). *Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.0)*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. URL: http://hdl.handle.net/11372/LRT-2282.

Schneider, Nathan, Emily Danchik, et al. (2014). "Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut". In: *Transactions of the Association for Computational Linguistics* 2, pp. 193–206. URL: https://www.cs.washington.edu/publications/discriminative-lexical-semantic-segmentation-gaps-running-mwe-gamut.

Schneider, Nathan and Noah A. Smith (2015). "A Corpus and Model Integrating Multiword Expressions and Supersenses". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. URL: https://www.aclweb.org/anthology/N15-1177.

Schulte im Walde, Sabine, Stefan Müller, and Stefan Roller (2013). "Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds". In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pp. 255–265. URL: https://www.aclweb.org/anthology/S13-1038.

Sheinfux, Livnat Herzig et al. (2019). "Verbal Multiword Expressions: Idiomaticity and flexibility". In: *Representation and parsing of multiword expressions: Current trends*. Ed. by Yannick Parmentier and Jakub Waszczuk. Language Science Press, pp. 35–68. URL: https://langsci-press.org/catalog/book/202.

Sidtis, Diana Van Lancker (2009). "Formulaic and novel language in a 'dual process' model of language competence: Evidence from surveys, speech samples, and schemata". In: *Typological Studies in Language*. John Benjamins Publishing Company. URL: https://doi.org/10.1075/tsl.83.11van.

Siyanova-Chanturia, Anna (2013). "Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and finding". In: *Mental Lexicon*, pp. 245–268. URL: https://doi.org/10.1075/ml.8.2.06siy.

Sommerauer, Pia and Antske Fokkens (2019). "Conceptual Change and Distributional Semantic Models: an Exploratory Study on Pitfalls and Possibilities". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pp. 223–233. DOI: 10.18653/v1/W19-4728. URL: https://www.aclweb.org/anthology/W19-4728.

Sporleder, Caroline and Linlin Li (2009). "Unsupervised Recognition of Literal and Non-Literal Use of Idiomatic Expressions". In: *Proceedings of the 12th Conference of the European Chapter of the ACL*. Association for Computational Linguistics, pp. 754–762. URL: https://www.aclweb.org/anthology/E09-1086.

Tahmasebi, Nina, Lars Borin, and Adam Jatowt (2018). "Survey of Computational Approaches to Diachronic Conceptual Change". In: *CoRR* abs/1811.06278. arXiv: 1811.06278. URL: http://arxiv.org/abs/1811.06278.

Tahmasebi, Nina and Thomas Risse (2017). "Finding Individual Word Sense Changes and their Delay in Appearance". In: *Proceedings of Recent Advances in Natural Language Processing*, pp. 741–749. URL: https://aclbg.org/proceedings/2017/RANLP%202017/pdf/RANLP095.pdf.

Tiedemann, Jörg (2018). "Emerging Language Spaces Learned From Massively Multilingual Corpora". In: *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*, pp. 188–197. URL: http://ceur-ws.org/Vol-2084/shortplus4.pdf.

University of Helsinki, The Department of Finnish, Finno-Ugrian and Scandinavian Studies, Institute for the Languages of Finland, and Heikki Paunonen (2014). *The Longitudinal Corpus of Finnish Spoken in Helsinki (1970s, 1990s and 2010s)*. URL: http://urn.fi/urn:nbn:fi:lb-2014073041.

Vecchi, Eva M. et al. (2016). "Spicy Adjectives and Nominal Donkeys: Capturing Semantic Deviance Using Compositionality in Distributional Spaces". In: *Cognitive Science* 41.1, pp. 102–136. DOI: 10.1111/cogs.12330. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12330.

Virpioja, Sami et al. (2013). *Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline*. Tech. rep. Aalto University. URL: https://morfessor.readthedocs.io/en/latest/general.html.

Wahl, Alexander and Stefan Th. Gries (2018). "Multi-word Expressions: A Novel Computational Approach to Their Bottom-Up Statistical Extraction". In: *Lexical Collocation Analysis: Advances and Applications*. Ed. by Pascual Cantos-Gómez and Moisés Almela-Sánchez. Cham: Springer International Publishing, pp. 85–109. ISBN: 978-3-319-92582-0. DOI: 10.1007/978-3-319-92582-0_5. URL: https://doi.org/10.1007/978-3-319-92582-0_5.

Wulff, Stefanie (2013). "Words and Idioms". In: *The Oxford Handbook of Construction Grammar*. Oxford University Press, pp. 274–289.

Yaneva, Victoria et al. (2017). "Cognitive Processing of Multiword Expressions in Native and Non-native Speakers of English: Evidence from Gaze Data". In: *Computational and Corpus-Based Phraseology, Second International Conference, Europhras 2017 London, UK, November 13–14, 2017, Proceedings*, pp. 363–379. DOI: 10.1007/978-3-319-69805-2_26. URL: https://www.researchgate.net/publication/320658248_Cognitive_Processing_of_Multiword_Expressions_in_Native_and_Non-native_Speakers_of_English_Evidence_from_Gaze_Data.

Yao, Zijun et al. (2018). "Dynamic Word Embeddings for Evolving Semantic Discovery". In: *WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining*. URL: https://doi.org/10.1145/3159652.3159703.

Zhong, Zhi and Hwee Tou Ng (2010). "It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text". In: *Proceedings of the ACL 2010 System Demonstrations*. Association for Computational Linguistics, pp. 78–83. URL: https://www.aclweb.org/anthology/P10-4014.

Zhu, Yi, Ivan Vulić, and Anna Korhonen (2019). "A Systematic Study of Leveraging Subword Information for Learning Word Representations". In: *Proceedings of NAACL-HLT 2019*. Association for Computational Linguistics, pp. 912–932. URL: https://www.aclweb.org/anthology/N19-1097.

Appendices

Appendix A Preprocessing Data Fixes

The process of fixing words was as follows:

1. If the word has w in it and does not have a voikko analysis, and
2. After doing the replacement w -> v, the corrected form was found in voikko
3. Then the word is corrected in the data.

Example corrected sentence:

> Kun hän noin järkähtämättä piti silmänsä kädessään olewassa kirjeessä, häntä woi luulla lewollisimmaksi lapseksi, jolla ei wielä ollut yhtään wakaisempaa ajatusta.

was fixed to

> Kun hän noin järkähtämättä piti silmänsä kädessään olevassa kirjeessä, häntä voi luulla **lewollisimmaksi** lapseksi, jolla ei vielä ollut yhtään vakaisempaa ajatusta.

Note the bolded word (a form of the adjective *levollinen* ('restful')), which was not found in the voikko vocabulary.

Appendix B Voikko Attribute Mapping

Mappings to English for voikko class names and Finnish noun case names[47] are shown in tables 33 and 34. Only the singular endings and the simplest forms are shown in the latter table[48].

Finnish	English
nimisana	noun
teonsana	verb
laatusana	adjective
nimisana_laatusana	noun_adjective
nimi	name
etunimi	firstname
sukunimi	lastname
paikannimi	location
lukusana	numeral
kieltosana	negation
lyhenne	abbreviation
seikkasana	adverb
asemosana	pronoun
etuliite	prefix
suhdesana	pre/postposition
sidesana	conjunction
huudahdussana	exclamation

Table 33: Voikko class name mapping

[47] Mappings copied from https://github.com/fergusq/tampio/blob/master/inflect.py.

[48] List taken from https://en.wikipedia.org/wiki/Finnish_noun_cases.

Finnish	English	Suffix(es)	English
nimento	nominative	-	
omanto	genitive	-n	of, 's
osanto	partitive	-a, -ä	(object)
olento	essive	-na, -nä	as
tulento	translative	-ksi	into
sisaolento	inessive	-ssa, -ssä	in
sisaeronto	elative	-sta, -stä	from
sisatulento	illative	-an, -en, etc	into
ulkoolento	adessive	-lla, -llä	at, on
ulkoeronto	ablative	-lta, -ltä	from
ulkotulento	allative	-lle	to
vajanto	abessive	-tta, -ttä	without
keinonto	instructive	-n	with, using
seuranto	comitative	-ne-	together
kerrontosti	adverb	-sti	-

Table 34: Voikko case name mapping